

Volume 13 Issue 1

January 2022



ISSN 2156-5570(Online)

ISSN 2158-107X(Print)

Editorial Preface

From the Desk of Managing Editor...

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Kohei Arai
Editor-in-Chief
IJACSA
Volume 13 Issue 1 January 2022
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Editor-in-Chief

Dr. Kohei Arai - Saga University

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

Associate Editors

Alaa Sheta

Southern Connecticut State University

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

Domenico Ciuonzo

University of Naples, Federico II, Italy

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

Dorota Kaminska

Lodz University of Technology

Domain of Research: Artificial Intelligence, Virtual Reality

Elena Scutelnicu

"Dunarea de Jos" University of Galati

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

In Soo Lee

Kyungpook National University

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

Krassen Stefanov

Professor at Sofia University St. Kliment Ohridski

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

Renato De Leone

Università di Camerino

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

Xiao-Zhi Gao

University of Eastern Finland

Domain of Research: Artificial Intelligence, Genetic Algorithms

CONTENTS

Paper 1: Performance Impact of Type-I Virtualization on a NewSQL Relational Database Management System

Authors: J. Bryan Osborne

PAGE 1 – 8

Paper 2: Knock Knock, Who's There: Facial Recognition using CNN-based Classifiers

Authors: Qiyu Sun, Alexander Redej

PAGE 9 – 16

Paper 3: Design of Smart IoT Device for Monitoring Short-term Exposure to Air Pollution Peaks

Authors: Eric Nizeyimana, Jimmy Nsenga, Ryosuke Shibasaki, Damien Hanyurwimfura, JunSeok Hwang

PAGE 17 – 24

Paper 4: Robust Facial Recognition System using One Shot Multispectral Filter Array Acquisition System

Authors: M. Eléonore Elvire HOUSSOU, A. Tidjani SANDA MAHAMA, Pierre GOUTON, Guy DEGLA

PAGE 25 – 33

Paper 5: Detecting Distributed Denial of Service in Network Traffic with Deep Learning

Authors: Muhammad Rusyaidi, Sardar Jaf, Zunaidi Ibrahim

PAGE 34 – 41

Paper 6: Proficient Networking Protocol for BPLC Network Built on Adaptive Multicast, PNP-BPLC

Authors: Ali Md Liton, Zhi Ren, Dong Ren, Xin Su

PAGE 42 – 48

Paper 7: Method for Improvement of Ocean Wind Speed Estimation Accuracy by Taking into Account the Relation between Wind Speed and Wind Direction

Authors: Kohei Arai, Kenta Azuma

PAGE 49 – 57

Paper 8: A Study of Security Impacts and Cryptographic Techniques in Cloud-based e-Learning Technologies

Authors: Lavanya-Nehan Degambur, Sheeba Armoogum, Sameerchand Pudaruth

PAGE 58 – 66

Paper 9: Various Antenna Structures Performance Analysis based Fuzzy Logic Functions

Authors: Chafaa Hamrouni, Aarif Alutaybi, Slim Chaoui

PAGE 67 – 71

Paper 10: Empirical Analysis Measuring the Performance of Multi-threading in Parallel Merge Sort

Authors: Muhyidean Altarawneh, Umur Inan, Basima Elshqairat

PAGE 72 – 78

Paper 11: Special Negative Database (SNDB) for Protecting Privacy in Big Data

Authors: Tamer Abdel Latif Ali, Mohamed Helmy Khafagy, Mohamed Hassan Farrag

PAGE 79 – 91

Paper 12: Drug Sentiment Analysis using Machine Learning Classifiers

Authors: Mohammed Nazim Uddin, Md. Ferdous Bin Hafiz, Sohrab Hossain, Shah Mohammad Mominul Islam

PAGE 92 – 100

Paper 13: Effective Malware Detection using Shapely Boosting Algorithm

Authors: Rajesh Kumar, Geetha S

PAGE 101 – 111

Paper 14: Using a Rule-based Model to Detect Arabic Fake News Propagation during Covid-19

Authors: Fatimah L. Alotaibi, Muna M. Alhammad

PAGE 112 – 119

Paper 15: Hybrid Deep Neural Network Model for Detection of Security Attacks in IoT Enabled Environment

Authors: Amit Sagu, Nasib Singh Gill, Preeti Gulia

PAGE 120 – 127

Paper 16: A Node Monitoring Agent based Handover Mechanism for Effective Communication in Cloud-Assisted MANETs in 5G

Authors: B.V.S Uma Prathyusha, K.Ramesh Babu

PAGE 128 – 136

Paper 17: Design of an Intelligent Hydroponics System to Identify Macronutrient Deficiencies in Chili

Authors: Deffa Rahadiyan, Sri Hartati, Wahyono, Andri Prima Nugroho

PAGE 137 – 145

Paper 18: Automatic Fake News Detection based on Deep Learning, FastText and News Title

Authors: Youssef Taher, Adelmoutalib Moussaoui, Fouad Moussaoui

PAGE 146 – 158

Paper 19: Determine the Level of Concentration of Students in Real Time from their Facial Expressions

Authors: Bouhlal Meriem, Habib Benlahmar, Mohamed Amine Naji, Elfilali Sanaa, Kaiss Wijdane

PAGE 159 – 166

Paper 20: Medical Image Cryptanalysis using Adaptive, Lightweight Neural Network based Algorithm for IoT based Secured Cloud Storage

Authors: M V Narayana, Ch Subba Lakshmi, Rishi Sayal

PAGE 167 – 173

Paper 21: Analysis of Logistics Service Quality and Customer Satisfaction during COVID-19 Pandemic in Saudi Arabia

Authors: Amjaad Bahamdain, Zahyah H. Alharbi, Muna M. Alhammad, Tahani Alqurashi

PAGE 174 – 180

Paper 22: Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer

Authors: Abdelrahman Elsharif Karrar

PAGE 181 – 188

Paper 23: Dynamic Deployment of Road Side Units for Reliable Connectivity in Internet of Vehicles

Authors: Abdulwahab Ali Almazroi, Muhammad Ahsan Qureshi

PAGE 189 – 194

Paper 24: Customer Satisfaction with Digital Wallet Services: An Analysis of Security Factors

Authors: Dewan Ahmed Muhtasim, Siok Yee Tan, Md Arif Hassan, Monirul Islam Pavel, Samiha Susmit

PAGE 195 – 206

Paper 25: Towards a Strategic IT GRC Framework for Healthcare Organizations

Authors: Fawaz Alharbi, Mohammed Nour A. Sabra, Nawaf Alharbe, Abdulrahman A. Almajed

PAGE 207 – 213

Paper 26: Ambulatory Monitoring of Maternal and Fetal using Deep Convolution Generative Adversarial Network for Smart Health Care IoT System

Authors: S. Venkatasubramanian

PAGE 214 – 222

Paper 27: Periapical Radiograph Texture Features for Osteoporosis Detection using Deep Convolutional Neural Network

Authors: Khasnur Hidjah, Agus Harjoko, Moh. Edi Wibowo, Rurie Ratna Shantiningih

PAGE 223 – 232

Paper 28: Prediction of Diabetic Obese Patients using Fuzzy KNN Classifier based on Expectation Maximization, PCA and SMOTE Algorithms

Authors: Ibrahim Eldesouky Fattoh, Soha Safwat

PAGE 233 – 238

Paper 29: Evaluation Optimal Prediction Performance of MLMs on High-volatile Financial Market Data

Authors: Yao HongXing, Hafiz Muhammad Naveed, Muhammad Usman Answer, Bilal Ahmed Memon, Muhammad Akhtar

PAGE 239 – 246

Paper 30: Optimize and Secure Routing Protocol for Multi-hop Wireless Network

Authors: Salwa Othmen, Wahida Mansouri, Somia Asklany, Wided Ben Daoud

PAGE 247 – 253

Paper 31: A Machine Learning Approach to Weather Prediction in Wireless Sensor Networks

Authors: Suvarna S Patil, B.M.Vidyavathi

PAGE 254 – 259

Paper 32: Detecting Diabetic Retinopathy in Fundus Images using Combined Enhanced Green and Value Planes (CEGVP) with k-NN

Authors: Minal Hardas, Sumit Mathur, Anand Bhaskar

PAGE 260 – 268

Paper 33: A Novel Secure Transposition Cipher Technique using Arbitrary Zigzag Patterns

Authors: Basil Al-Kasasbeh

PAGE 269 – 276

Paper 34: Feature Concatenation based Multilayered Sparse Tensor for Debond Detection Optical Thermography

Authors: Junaid Ahmed, Abdul Baseer, Guiyun Tian, Gulsher Baloch, Ahmed Ali Shah

PAGE 277 – 282

Paper 35: Unmoderated Remote Usability Testing: An Approach during Covid-19 Pandemic

Authors: Ambar Relawati, Guntur Maulana Zamroni, Yanuar Primanda

PAGE 283 – 289

Paper 36: 4PCDT: A Quantifiable Parameter-based Framework for Academic Software Project Management

Authors: Vikas S. Chomal, Jatinderkumar R. Saini, Hema Gaikwad, Ketan Kotecha

PAGE 290 – 297

Paper 37: Neuromarketing Solutions based on EEG Signal Analysis using Machine Learning

Authors: Asad Ullah, Gulsher Baloch, Ahmed Ali, Abdul Baseer Buriro, Junaid Ahmed, Bilal Ahmed, Saba Akhtar

PAGE 298 – 304

Paper 38: Tomato Leaf Disease Detection using Deep Learning Techniques

Authors: Nagamani H S, Sarojadevi H

PAGE 305 – 311

Paper 39: Feature Selection Pipeline based on Hybrid Optimization Approach with Aggregated Medical Data

Authors: Palwinder Kaur, Rajesh Kumar Singh

PAGE 312 – 320

Paper 40: Educational Data Mining to Identify the Patterns of Use made by the University Professors of the Moodle Platform

Authors: Johan Calderon-Valenzuela, Keisi Payihuanca-Mamani, Norka Bedregal-Alpaca

PAGE 321 – 328

Paper 41: Assessing the Quality of Educational Websites in Sudan using Quality Model Criteria through an Electronic Tool

Authors: Asim Seedahmed Ali Osman

PAGE 329 – 334

Paper 42: Secure Inter-Domain Routing for Resisting Unknown Attacker in Internet-of-Things

Authors: Bhavana A, Nanda Kumar A N

PAGE 335 – 342

Paper 43: Snowball Framework for Web Service Composition in SOA Applications

Authors: Mohamed Elkholly, Youcef Bagdadi, Marwa Marzouk

PAGE 343 – 350

Paper 44: Selection of Requirement Elicitation Techniques: A Neural Network based Approach

Authors: Mohd Muqem, Sultan Ahmad, Jabeen Nazeer, Md. Faizan Farooqui, Afroj Alam

PAGE 351 – 359

Paper 45: The Performance of Personality-based Recommender System for Fashion with Demographic Data-based Personality Prediction

Authors: Iman Paryudi, Ahmad Ashari, Khabib Mustofa

PAGE 360 – 368

Paper 46: A Greedy-based Algorithm in Optimizing Student's Recommended Timetable Generator with Semester Planner

Authors: Khyrina Airin Fariza Abu Samah, Siti Qamalia Thusree, Ahmad Firdaus Ahmad Fadzil, Lala Septem Riza, Shafaf Ibrahim, Noraini Hasan

PAGE 369 – 375

Paper 47: Development of an Efficient Electricity Consumption Prediction Model using Machine Learning Techniques

Authors: Ghaidaa Hamad Alraddadi, Mohamed Tahar Ben Othman

PAGE 376 – 384

Paper 48: Critical Review of Technology-Enhanced Learning using Automatic Content Analysis

Authors: Amalia Rahmah, Harry B. Santoso, Zainal A. Hasibuan

PAGE 385 – 394

Paper 49: A Knowledge-based Expert System for Supporting Security in Software Engineering Projects

Authors: Ahmad Azzazi, Mohammad Shkoukani

PAGE 395 – 400

Paper 50: Performance Comparison between Lab-VIEW and MATLAB on Feature Matching-based Speech Recognition System

Authors: Edita Rosana Widasari, Barlian Henryranu Prasetio, Dian Eka Ratnawati

PAGE 401 – 407

Paper 51: A New Priority Rule for Initial Ordering of Jobs in Permutation Flowshop Scheduling Problems

Authors: B. Dhanasakkaravarthi, A. Krishnamoorthy

PAGE 408 – 415

Paper 52: Preserving Location Privacy in the IoT against Advanced Attacks using Deep Learning

Authors: Abdullah S. Alyousef, Karthik Srinivasan, Mohamad Shady Alrahal, Majdah Alshammari, Mousa Al-Akhras

PAGE 416 – 427

Paper 53: A Conceptual User Experience Evaluation Model on Online Systems

Authors: Norhanisha Yusof, Nor Laily Hashim, Azham Hussain

PAGE 428 – 438

Paper 54: Applying Artificial Intelligence in Retrieving Design Solution

Authors: Y. Moubachir, B. Hamri, S. Taibi

PAGE 439 – 444

Paper 55: State-of-the-Art Approach to e-Learning with Cutting Edge NLP Transformers: Implementing Text Summarization, Question and Distractor Generation, Question Answering

Authors: Spandan Patil, Lokshana Chavan, Janhvi Mukane, Deepali Vora, Vidya Chitre

PAGE 445 – 453

Paper 56: A Regression Model to Predict Key Performance Indicators in Higher Education Enrollments

Authors: Ashraf Abdelhadi, Suhaila Zainudin, Nor Samsiah Sani

PAGE 454 – 460

Paper 57: A Novel Stance based Sampling for Imbalanced Data

Authors: Isha Agarwal, Dipti Rana, Aemie Jariwala, Sahil Bondre

PAGE 461 – 467

Paper 58: Energy Efficient and Quality-of-Service Aware Routing using Underwater Wireless Sensor Networks

Authors: P. Sathya, P. Sengottuvelan

PAGE 468 – 474

Paper 59: The Trend of Segmentation for Arabic Handwritten Touching Characters

Authors: Ahmed Mansoor Mohsen Algaradi, Mohd Sanusi Azmi, Intan Ermahani A. Jalil, Abdulwahab Fuad Ayyash Hashim, Afrah Abdullah Muhammad Al-Malki

PAGE 475 – 479

Paper 60: What Influences Customer's Trust on Online Social Network Sites (SNSs) Sellers?

Authors: Ramona Ramli, Asmidar Abu Bakar, Fiza Abdul Rahim

PAGE 480 – 489

Paper 61: New Textual Authentication Method to Resistant Shoulder-Surfing Attack

Authors: Islam Abdalla Mohamed Abass, Loay F.Hussein, Tarak kallel, Anis Ben Aissa

PAGE 490 – 496

Paper 62: CovSeg-Unet: End-to-End Method-based Computer-Aided Decision Support System in Lung COVID-19 Detection on CT Images

Authors: Fatima Zahra EL BIACH, Imad IALA, Hicham LAANAYA, Khalid MINAOUI

PAGE 497 – 504

Paper 63: Elevint: A Cloud-based Internet of Elevators

Authors: Sarah Mohammed Aljadani, Shahd Mohammed Almutairi, Saja Saeed Ghaleb, Lama Al Khuzayem

PAGE 505 – 513

Paper 64: Moving Object Detection over Wireless Visual Sensor Networks using Spectral Dual Mode Background Subtraction

Authors: Ahmed M. AbdelTawab, M.B. Abdelhalim, S.E.D. Habib

PAGE 514 – 523

Paper 65: Enhancing the Security of Digital Image Encryption using Diagonalize Multidimensional Nonlinear Chaotic System

Authors: Mahmoud I. Moussa, Eman I. Abd El-Latif, Nawaz Majid

PAGE 524 – 533

Paper 66: A Visual-Range Cloud Cover Image Dataset for Deep Learning Models

Authors: Muhammad Umair, Manzoor Ahmed Hashmani

PAGE 534 – 541

Paper 67: Blockchain in the Quantum World

Authors: Arman Rasoodl Faridi, Faraz Masood, Ali Haider Thabet Shamsan, Mohammad Luqman, Monir Yahya Salmony

PAGE 542 – 552

Paper 68: Design and Implementation of Deep Depth Decision Algorithm for Complexity Reduction in High Efficiency Video Coding (HEVC)

Authors: Helen K Joy, Manjunath R Kounte, B K Sujatha

PAGE 553 – 560

Paper 69: The Pragmatics of Function Words in Fiction

Authors: Ayman Farid Khafaga

PAGE 561 – 570

Paper 70: Fusion of BIFFOA and Adaptive Two-Phase Mutation for Helmetless Motorcyclist Detection

Authors: Sutikno, Agus Harjoko, Afiahayati

PAGE 571 – 581

Paper 71: AI-based System for the Detection and Prevention of COVID-19

Authors: Sofien Chokri, Wided Ben Daoud, Wasma Hanini, Sami Mahfoudhi, Amel Makhoulouf

PAGE 582 – 591

Paper 72: Human Emotion Recognition by Integrating Facial and Speech Features: An Implementation of Multimodal Framework using CNN

Authors: P V V S Srinivas, Pragnyaban Mishra

PAGE 592 – 603

Paper 73: BERT based Named Entity Recognition for Automated Hadith Narrator Identification

Authors: Emha Taufiq Luthfi, Zeratul Izzah Mohd Yusoh, Burhanuddin Mohd Aboobaider

PAGE 604 – 611

Paper 74: An Early Intervention Technique for At-Risk Prediction of Higher Education Students in Cloud-based Virtual Learning Environment using Classification Algorithms during COVID-19

Authors: Arul Leena Rose.P.J, Ananthi Claral Mary.T

PAGE 612 – 621

Paper 75: Balanced Schedule on Storm for Performance Enhancement

Authors: Arwa Z. Selim, Noha E. El-Affar, I. M. Hanafy, Wael A. Awad

PAGE 622 – 632

Paper 76: Extract Concept using Subtitles in MOOC

Authors: Aarika Kawtar, Habib Benlahmar, Mohamed Amine Naji, Elfilali Sanaa, Zouheir Banou

PAGE 633 – 638

Paper 77: Identification of Coronary Heart Disease through Iris using Gray Level Co-occurrence Matrix and Support Vector Machine Classification

Authors: Vincentius Abdi Gunawan, Leonardus Sandy Ade Putra, Fitri Imansyah, Eka Kusumawardhani

PAGE 639 – 648

Paper 78: Performance of Data Reduction Algorithms for Wireless Sensor Network (WSN) using Different Real-Time Datasets: Analysis Study

Authors: M. K. Hussein, Ion Marghescu, Nayef.A.M. Alduais

PAGE 649 – 661

Paper 79: A Global Survey of Technological Resources and Datasets on COVID-19

Authors: Manoj Muniswamaiah, Tilak Agerwala, Charles C. Tappert

PAGE 662 – 687

Paper 80: A Comparison between Online and Offline Health Seeking Information using Social Networks for Patients with Chronic Health Conditions

Authors: Andrew Kear, Simon Talbot

PAGE 688 – 699

Paper 81: Predicting Cyber-Attack using Cyber Situational Awareness: The Case of Independent Power Producers (IPPs)

Authors: Akweley Henry Matey, Paul Danquah, Godfred Yaw Koi-Akrofi

PAGE 700 – 709

Paper 82: Design and Performance Analysis of Anti-Surge Control Mechanism for Compressor System using Neural Networks

Authors: Divya M.N, Narayanappa C.K, S L Gangadhariah, V Nuthan Prasad

PAGE 710 – 718

Paper 83: Design and Implementation of True Parallelism Quad-Engine Cybersecurity Architecture on FPGA

Authors: Nada Qaim Mohammed, Amiza Amir, Muataz Hammed Salih, Badlishah Ahmad

PAGE 719 – 724

Paper 84: Cotton Crop Yield Prediction using Data Mining Technique

Authors: Amiksha Ashok Patel, Dhaval Kathiriya

PAGE 725 – 731

Paper 85: Data Analysis of Coronavirus CoVID-19: Study of Spread and Vaccination in European Countries

Authors: Hela Turki, Kais Khrouf

PAGE 732 – 737

Paper 86: Design of Low Cost Bio-impedance Measuring Instrument

Authors: Rajesh Birok, Rajiv Kapoor

PAGE 738 – 749

Paper 87: Detecting Irony in Arabic Microblogs using Deep Convolutional Neural Networks

Authors: Linah Alhaidari, Khaled Alyoubi, Fahd Alotaibi

PAGE 750 – 756

Paper 88: Analysis about Benefits of Software-Defined Wide Area Network: A New Alternative for WAN Connectivity

Authors: Catherine Janir´e Mena Diaz, Laberiano Andrade-Arenas, Javier Gustavo Utrilla Arellano, Miguel Angel Cano Lengua

PAGE 757 – 767

Paper 89: Data Recovery Approach for Fault-Tolerant IoT Node

Authors: Perigisetty Vedavalli, Deepak Ch

PAGE 768 – 774

Paper 90: An Enhanced Traffic Split Routing Heuristic for Layer 2 and Layer 1 Services

Authors: Ahlem Harchay, Abdelwahed Berguiga, Ayman Massaoudi

PAGE 775 – 781

Paper 91: AuSDiDe: Towards a New Authentication System for Distributed and Decentralized Structure based on Shamir's Secret Sharing

Authors: Omar SEFRAOUI, Afaf Bouzidi, Kamal Ghoumid, El Miloud Ar-Reyouchi

PAGE 782 – 787

Paper 92: Keyphrases Concentrated Area Identification from Academic Articles as Feature of Keyphrase Extraction: A New Unsupervised Approach

Authors: Mohammad Badrul Alam Miah, Suryanti Awang, Md. Saiful Azad, Md Mustafizur Rahman

PAGE 788 – 796

Paper 93: Transfer Learning based Performance Comparison of the Pre-Trained Deep Neural Networks

Authors: Jayapalan Senthil Kumar, Syahid Anuar, Noor Hafizah Hassan

PAGE 797 – 805

Paper 94: Augmented Reality: Prototype for the Teaching-Learning Process in Peru

Authors: Shalom Adonai Huaraz Morales, Laberiano Andrade-Arenas, Alexi Delgado, Enrique Lee Huamani

PAGE 806 – 815

Paper 95: On the Long Tail Products Recommendation using Tripartite Graph

Authors: Arlisa Yuliawati, Hamim Tohari, Rahmad Mahendra, Indra Budi

PAGE 816 – 822

Paper 96: Machine Learning Applied to Prevention and Mental Health Care in Peru

Authors: Edwin Kcomf Ponce, Melissa Flores Cruz, Laberiano Andrade-Arenas

PAGE 823 – 831

Paper 97: Modeling and Predicting Blood Flow Characteristics through Double Stenosed Artery from Computational Fluid Dynamics Simulations using Deep Learning Models

Authors: Ishaq Raihan Jamil, Mayeesha Humaira

PAGE 832 – 841

Paper 98: NLI-GSC: A Natural Language Interface for Generating SourceCode

Authors: Aaqib Ahmed R.H. Ansari, Deepali R. Vora

PAGE 842 – 853

Paper 99: Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic

Authors: Fatima Alhaj, Ali Al-Haj, Ahmad Sharieh, Riad Jabri

PAGE 854 – 860

Paper 100: Investigation Framework for Cloud Forensics using Dynamic Genetic-based Clustering

Authors: Mohammed Y. Alkhanafseh, Mohammad Qataweh, Wesam Almobaideen

PAGE 861 – 873

Paper 101: Towards a Low-Cost FPGA Micro-Server for Big Data Processing

Authors: Mohamed Abouzahir, Khalifa Elmansouri, Rachid Latif, Mustapha Ramzi

PAGE 874 – 884

Paper 102: Assessing and Proposing Countermeasures for Cyber-Security Attacks

Authors: Ali Al-Zahrani

PAGE 885 – 895

Paper 103: ASM-ROBOT: A Cyber-Physical Home Automation Controller with Memristive Reconfigurable State Machine

Authors: Kennedy Chinedu Okafor, Omowunmi Mary Longe

PAGE 896 – 912

Performance Impact of Type-I Virtualization on a NewSQL Relational Database Management System

J. Bryan Osborne

Computing & Mathematics
Oral Roberts University, Tulsa, OK, USA

Abstract—For more than 40 years, the relational database management system (RDBMS) and the atomicity, consistency, isolation, durability (ACID) transaction guarantees provided through its use have been the standard for data storage. The advent of Big Data created a need for new storage approaches that led to NoSQL technologies, which rely on basic availability, soft-state, eventual consistency (BASE) transactions. Over the last decade, NewSQL RDBMS technology has emerged, providing the benefits of RDBMS ACID transaction guarantees and the performance and scalability of NoSQL databases. The reliance on virtualization in IT has continued to grow, but an investigation of current academic literature identified a void regarding the performance impact of virtualization of NewSQL databases. To help address the lack of research in this area, a quantitative experimental study was designed and carried out to answer the central research question, "What is the performance impact of Type-I virtualization on a NewSQL RDBMS?" VMware ESXi virtualization software, NuoDB RDBMS, and OLTP-Bench software were used to execute a mixed-load benchmark. Performance metrics were collected comparing bare metal and virtualized environments, and the data analyzed statistically to evaluate five hypotheses related to CPU utilization, memory utilization, disk and network input-output (I/O) rates, and database transactions per second. Findings indicated a negative performance impact on CPU and memory utilization, as well as network I/O rates. Performance improvements were noted in disk I/O rates and database transactions-per-second.

Keywords—Database benchmarking; NewSQL; relational database; virtualization

I. INTRODUCTION

Drastically changing paradigms of data management and storage, Big Data continues to be a disruptive technology in the world of computing [1]. To provide acceptable performance, databases required new architectures and a built-from-scratch approach to overcome performance limitations inherent to traditional relational database management systems [RDBMSs; 2]. The newest of these databases are referred to as NewSQL and have been gaining traction with their ability to support massively large datasets, provide atomicity, consistency, isolation, durability (ACID) transaction guarantees, support for structured query language (SQL) queries, and do so with the performance provided by NoSQL solutions [3].

The use of virtualization in the modern era provides information technology (IT) management the ability to more efficiently manage resources through improved utilization rates and allows for greater flexibility of existing equipment and increased scalability of physical servers [4]. Pogarcic, Krnjak and Ozanic [5] demonstrated that virtualization could provide

decreased capital and operation costs for businesses in areas of actual and procurement costs for server hardware, utility, and administration costs. Virtualization technology provides the foundation for cloud computing, which continues to shape the world of modern IT [6]. Cloud computing providers regularly utilize virtualization's flexibility and scalability to maximize revenue and meet the increasing demands for their services [7]. Cloud computing relies heavily on virtualization to provide server and database services to customers [8].

NewSQL databases have demonstrated superiority in performance testing against NoSQL databases involving Internet of Things (IoT) applications [9] and as a solution for Big Data OLTP applications [10]. Performance comparisons of NewSQL offerings appear in the literature but are limited to comparisons of NewSQL databases against each other in bare metal [9, 11], virtualized [12], and cloud [13] environments. There is a lack of published research on the performance implications of the virtualization of NewSQL systems. The continued increase in the use of virtualization technology in data centers and the increase in the implementation of NewSQL database systems in cloud computing indicated a need for this research effort.

The strong presence of virtualization as a technology in IT services including the cloud [4], the continued growth in the use of NewSQL databases [3], and the simultaneous use of both, presented a need to understand the impact of one on the other. A review of extant literature indicated a void with respect to the examination and quantification of the impact of virtualization on NewSQL RDBMS. Therefore, the research effort posed a central question of "What is the performance impact of Type-I virtualization on a NewSQL RDBMS?" NewSQL databases are relational by nature and therefore the measures related to RDBMS are relevant, as is the impact of virtualization on relevant system metrics. The use of throughput of RDBMS software, measured by transactions-per-second (TPS) as suggested by Bitton, et al. [14], remains a prevalent and accepted metric for performance along with the use of benchmarking software to perform testing [15]. Therefore, performance measures of system CPU, memory, disk and network I/O, and NewSQL database throughput were utilized as the dependent variables in this research. The impact on these dependent variables due to changes in the independent variable, stated as the condition of the system being bare metal or Type-I virtualized, led to the central question being operationalized into five null (and corresponding alternative) hypotheses, one for each of the five performance measures to be tested. The research quantified the level of impact virtualization caused to a system running a NewSQL database

via an experimental design constructed to measure and analyze variables relevant to system performance in both bare metal and virtualized environments. The investigation contributes to the body of knowledge by filling a void in the academic literature regarding the performance impact of virtualization of NewSQL databases. Further, it provides information relevant to potentially needed modifications to a system to more efficiently host a NewSQL RDBMS.

The rest of the paper is organized as follows. Section II briefly discusses virtualization, NewSQL database management systems, and research relevant to the project. Section III describes the experimental methodology, hardware and software configuration, benchmarking software used in the test environment and statistical methods utilized. Section IV contains the results of the statistical analysis of the data collected and a discussion of the results of the analysis. Finally, Section V presents conclusions reached.

II. BACKGROUND AND RELATED WORK

A. Virtualization

Type I hypervisors interface directly with the computer's underlying hardware on which they run and map the physical resources of a computer to the VM being hosted, as shown in Fig. 1. This architectural approach has led to these types of hypervisors being referred to as native [16] or bare metal [17] hypervisors as they execute directly on the host computer and serve as a link between the virtual machines and the host computer.

The goal of virtualization is to use a layer of software to abstract the specific hardware of a computing machine from operating systems executing on the machine [18]. Thus, virtualization creates both an abstraction and encapsulation of the various components of the underlying hardware [19], and it is the role of the hypervisor to provide this abstraction as well as provide the virtual machine's integrity and isolation from other VMs on the same hardware [20]. Three major virtualization approaches, full virtualization, paravirtualization, and hardware-assisted virtualization, have been used to overcome the shortcomings of the x86 architecture [20].

The impact of virtualization caused by the insertion of a layer of processing between the guest OS and hardware has been demonstrated by extant research. In a review of 112 publications produced prior to 2016, Kao [21] found that the majority of evaluations of virtualized environments utilized benchmarks to measure performance. The specific benchmarks used by studies reviewed as part of this literature review mirrored those found by Kao [21] and reflected the emphasis found in the literature on using measurements of CPU, memory, disk, and network utilization as a basis for performance analysis. While it can be argued that differences exist between benchmarks and real-world workloads [22], benchmarking has a long history and provides for the gathering of metrics using a well-defined, universally accepted set of tests that can be compared across applications, operating systems, and hardware platforms [15]. Based on the review of all articles, the research focused on these metrics to assist in the evaluation of the impact of virtualization on NewSQL databases.

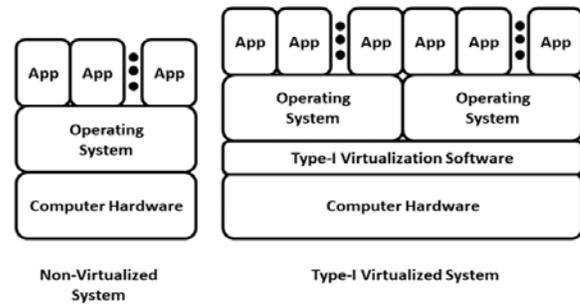


Fig. 1. Comparison of Non-virtualized and Type-1 Virtualized Systems.

B. NewSQL

The term "NewSQL" is generally attributed to a report issued by the 451 Group analysts Matthew Aslett, who is referring to what was at the time new vendor offerings of databases that supported SQL, ACID transactions, and the high performance and scalability of NoSQL databases [23]. A short time later, in a blog post on Communications of the ACM website, Michael Stonebraker argued that instead of continuing the "gold standard in enterprise computing" of multiple relational databases connected by extract-load-transform (ETL) processes, NewSQL database systems, with the characteristics listed above, address consistency issues and preserve SQL language capabilities [24]. In a blog posting, Stonebraker outlined five specific characteristics of NewSQL as having ACID transactional ability, concurrency control that was non-locking, high performance, possessed a distributed, horizontally scalable, shared-nothing architecture, and SQL as application language. Although these characteristics were not presented in the format of an academic paper, the qualifications listed are present in academic works discussing NewSQL systems [2, 10, 12, 25-27].

NewSQL systems can be categorized into three types, novel systems built from scratch utilizing a new architecture, middleware approaches, and Database-as-a-Service (DBaaS) offerings [2]. The research in this paper focuses on an example of the first category, which includes the characteristics outlined by Stonebraker [24] for NewSQL systems and is created through the writing of completely new code, thereby being free from architectural choices of existing systems [28]. The development of the system as a new application provides the ability to manage best the disk and memory storage, replication approaches, query optimization, and node communication, allowing better performance than systems built through the layering of existing technologies [2].

Research by Hrubaru and Fotache [29] evaluated the performance of RDBMS and NewSQL using the TPC-H benchmark, capturing and analyzing loading and query times, which are essentially throughput measures, as well as the amount of memory used by the DBMSs. Not surprisingly, the in-memory NewSQL instance used more memory than the two RDBMS but performed better in most cases in terms of loading and query times. Similarly, Oliveira and Bernardino [11] compared two NewSQL RDBMSs using TPC-H, comparing loading and query execution times, but an evaluation of memory utilization was not performed. In both of these

articles, as with Fatima and Wasnik [9], the experimental setup only used a single server to perform tests. Two studies involving performance measures of virtualized NewSQL environments were identified. Both focused on comparing RDBMS performances when running in virtual machines, but no evaluation was made comparing bare metal and virtual environments. Kaur and Sachdeva [12] compared throughput performance measures of four popular NewSQL databases but did not specify the exact tests performed, only providing latency times captured for read, write, and update operations and total execution time. Tests were performed using a single instance RDBMS running in a single Type II hypervisor. While relative performance measures provide some helpful information, it is limited due to the use of single-instance NewSQL servers and a Type II hypervisor, neither of which would be common in a larger scale production system. Borisenko and Badalyan [30] utilized a Type I hypervisor with individual cluster nodes running in separate VMs to evaluate two NewSQL RDBMSs. One of the NewSQL RDBMSs evaluated, Apache Ignite [31], would be considered by the definitions provided by Pavlo and Aslett [2] as a middleware approach. The other, VoltDB [32], is a new architecture, thereby providing a performance comparison between the two approaches. The workload utilized was described as "TPC-H like," and metrics gathered and analyzed consisted only of query execution times [30]. Although both investigations utilized virtualized environments, Borisenko and Badalyan [30] and Kaur and Sachdeva [12] treated the performance testing as if performed on bare metal and did not provide any insights into the performance impact of virtualization.

Given the relevance of cloud computing today and the importance of virtualization to cloud computing, the implementation of NewSQL databases in the cloud is very probable [25]. Previous research has identified comparative performance studies of NewSQL databases, but none that reflected the impact of Type-I virtualization on NewSQL databases. Further, the studies examined reflected the use of single-node systems, not reflecting the true distributed nature of NewSQL databases [2, 27, 28]. This paper fills the gap in the literature by providing a quantification of performance impacts to distributed NewSQL systems due to Type-I virtualization.

III. METHODOLOGY

The quantitative experimental research embodied in this study provides insight into the impact of Type-I virtualization on a NewSQL database, as well as a foundation for further exploration of potential mechanisms to best leverage virtualization in this new and important realm of computing. The research quantified the level of impact virtualization causes to a system running a NewSQL database via an experimental design constructed to measure and analyze variables relevant to system performance in both bare metal and virtualized environments. The central research question, "What is the performance impact of Type-I virtualization on NewSQL relational databases?" operationalized into five null hypotheses, each stating there was no difference in recorded values of the respective dependent variables during database benchmarking for a system hosting a NewSQL database

instance when comparing bare metal and Type-I virtualized environments.

The configuration of the system, bare metal or virtualized, reflects the independent variable in the experimental design. Metrics of CPU and memory utilization, disk and network I/O measurements, and the number of transactions-per-second executed by the databases represent dependent variables. These values were captured and recorded during benchmark tests executed on bare metal and virtualized systems built on identical servers. The values captured were then analyzed to quantify differences in performance between bare metal and virtualized systems.

The computer hardware used consisted of a Hewlett Packard Enterprise (HPE) Apollo r2600 chassis with 3 HPE dual-processor nodes. Each processor node was comprised of two Intel Xeon Broadwell E5-2698v4 2.4GHz CPUs, each with 14 cores and 50MB cache, 128GB of DDR4-2400 memory, and a single 480MB SATA solid-state drive (SSD). Internode communications are performed through 1 GbE Ethernet ports connected via an Omnipath 100Gb port. The computing nodes resided on an isolated network, ensuring only intended access to the machines.

The software configuration of each node varied based upon its functionality in the test. Each node ran the same operating system, CentOS 7.7.1908, an open source version of the Red Hat Enterprise Linux operating system [33]. Two of the three hardware nodes were designated as database cluster nodes using the recommended CentOS infrastructure server template, NuoDB NewSQL RDBMS, and performance metric collection software. The third node, configured to execute benchmark and load generation software, was configured as a developer's workstation to allow for the compiling of software with the GNU compiler.

In a recent study by Almassabi, Bawazeer [27], 13 of the top NewSQL databases were identified. From this group, NuoDB CE 4.0.4.2 [34] was selected for the current research effort as it meets the requirements defined by Stonebraker [24]. NuoDB is a distributed, peer-to-peer, ACID-compliant, elastic, and highly scalable relational database management system that falls into the "new architecture" class, elastic and highly scalable [35], and offers a fully functional, community edition version available at no cost.

NuoDB is designed to operate in bare metal, virtualized, and cloud environments [36]. The database management system has a two-tier, distributed architecture separating transactional and storage tiers. The transactional tier is an in-memory tier responsible for atomicity, consistency, and isolation aspects of the ACID transactional model and is designed to ensure fast access to data by applications. The storage management tier is responsible for the durability aspect, ensuring data is safely stored when committed, and providing data in case of a cache miss. Although all nodes are peers in a cluster, NuoDB nodes execute as either a Transaction Engine (TE) or a Storage Manager (SM). SMs maintain complete, consistent, independent copies of the entire database. TEs cache database tables in memory, accept database requests and execute SQL queries. For purposes of

this research, one of the two nodes functioned as an SM and the other database node as a TE. Installation of the NuoDB software was performed per the manufacturer's instructions [37]. Also installed on each of the database nodes was the performance metric gathering tool, "nmon" [38], a software tool written in the C programming language, demonstrated to be effective in capturing metrics in performance testing research [39, 40]. The nmon application captured data on database nodes for CPU utilization measured as a percent of total available, total system memory utilization measured in MB allocated, disk I/O measured in KB/s, and network I/O measured in KB/s, each of which represents a dependent variable in the experiment.

The third node ran the load-generating, benchmarking software, OLTP-Bench [41]. OLTP-Bench is an open-source, extensible, flexible benchmarking testbed capable of executing a number of existing benchmarks on both on-premise and cloud databases [42]. The Java source code for the OLTP-Bench software is open-source and available on GitHub and was downloaded and compiled for the CentOS platform on the benchmark node. The software was utilized to exercise the NewSQL database and system and provided the transactions-per-second measurement, a dependent variable, for all tests. Due to the different nature of online transaction processing (OLTP) and online analytical processing (OLAP), separate databases with different designs have typically been implemented [43]. A growing need for real-time analytics has generated a change in this paradigm, and new databases, like NewSQL, have been developed to provide support for these needs [44]. Subsequently, benchmarking tests to evaluate the performance of databases designed to handle mixed workloads are needed [45].

One such benchmark, currently supported by OLTP-Bench, is CH-benCHmark [46]. CH-benCHmark is a hybrid/mixed-workload benchmark designed to execute TPC-C (OLTP) and TPC-H-equivalent (OLAP) queries concurrently against a common set of database tables. The entities and relationships of the TPC-C model are implemented without modification, and only a slight modification to the TPC-H schema is made to ensure the integration into the TPC-C schema is non-intrusive. Previous work by Oliveira and Bernardino [11] and Hrubaru and Fotache [29] revealed issues running all 22 of the TPC-H queries against new architecture NewSQL databases due to lack of support for the SQL HAVING clause, view creation capabilities, and excessively long-running query conditions. Initial tests performed as part of this research revealed similar incompatibility issues as well as hang-ups in benchmark execution with the CH-benCHmark TPC-H comparable queries. The subset consisting of seven of the CH-benCHmark TPC-H comparable queries, which ran without issue, was placed in the OLTP-Bench configuration file to be used for testing.

Based on the assumption that a One-Way ANOVA would be utilized for statistical analysis, a sample size based upon a power analysis using the G*Power application [47] indicated a need for 21 bare metal and 21 virtualized runs of the benchmark test against the NewSQL databases. The first set of tests in the experiment in this research effort was the execution of the CH-benCHmark using the OLTP-Bench testbed against

the NuoDB database in a bare metal environment. The creation and loading of the test databases were performed in separate, sequential steps from the execution of the benchmark test to delineate the response of the system under test to these operations, as part of an effort to ensure "cold" runs of the benchmark workload [48]. The stopping and restarting of the database after loading causes a flushing of the database server buffer pool, eliminating data caching between runs and enhancing consistency in values of performance metrics. Preliminary benchmark runs revealed that system performance metrics stayed in a consistent range in benchmark runs lasting as long as one hour. It was also determined that the entire set of TPC-H like queries was completed in approximately five minutes, even in the presence of concurrent OLTP transactions. Therefore, benchmark runs of fifteen minutes were used for data collection runs to allow three sets of the TPC-H queries.

Once all experiments had been run under bare metal conditions, the "treatment" (virtualization software) was installed on the same hardware used for the first set of experiments. VMware 6.5 [49] served as the virtualization software utilized in the experiment. VMware is a major player in the virtualization space, garnering an 80.7 % share of the 2017 virtualization market [50]. A single virtual machine closely matching the specifications of the physical machine on which it resides was created on the nodes of the database cluster. Each database node used a thin-provisioned, 400GB disk configured using a SCSI-controller in dependent mode and located in a VMware datastore. The VMware Paravirtual SCSI controller was used following Dakic [51], Goldsand and Brown [52], and VMware [53]. The VMware VM Network was configured with the physical NIC connected to a vSwitch, which was connected to the virtual machine.

Each virtual machine was installed with the identical software configuration, i.e., operating system, NewSQL database, and performance metric gathering software, as was installed on the machine in its bare metal state. The same type and number of benchmark runs were performed, and performance data was collected. Once data collection was completed under virtualized conditions, the two sets of data required for the independent variable, bare metal versus virtualized environment, were made available for analysis.

IV. RESULTS AND DISCUSSION

Care was taken to ensure the appropriate statistical approach was taken in the comparison of data collected in bare metal and virtualized environments. Using SPSS, an evaluation to determine the normality of the individual datasets was completed via a Shapiro Wilk test, followed by either an ANOVA or Kruskal-Wallis H (K-W) test depending on whether a parametric or non-parametric test was needed to determine statistical significance. A summary of the inferential statistical tests and changes due to virtualization is provided in Table I, and a detailed description of the statistical results follows.

The results of the ANOVA analyzing CPU utilization for the SM node indicated that the difference in the means was statistically significant, $F(1,40)=3083.879$, $p<0.001$, and the null hypothesis was rejected. The rejection of the null hypothesis led to the acceptance of the alternative hypothesis

that there was a difference in CPU use when virtualization is implemented. A comparison of the mean values of CPU utilization between bare metal (M=1.492, SD=0.052) and virtualized (M=2.26, SD=0.057) environments for the SM node indicated an increase of 62.5%. In the case of the TE node, the ANOVA also demonstrated that the difference in the means is statistically significant, $F(1,40)=5027.822$, $p<0.001$, and the null hypothesis can be rejected. A comparison of the mean values of CPU utilization between bare metal (M=2.684, SD=0.049) and virtualized environments (M=4.546, SD=0.110) indicated an increase of 69.4 %.

TABLE I. RESULTS

Perf Measure	SM Virtualized vs. Bare Metal			TE Virtualized vs. Bare Metal		
	Test	p-value	Chg	Test	p-value	Chg
CPU Util	ANOVA	$p < 0.001$	62.5%	ANOVA	$p < 0.001$	69.4%
Memory Util	ANOVA	$p < 0.001$	1.4%	K-W	$p < 0.001$	2.7%
Disk I/O	K-W	$p < 0.001$	37.3%	K-W	$p < 0.001$	N/A
Network I/O	ANOVA	$p < 0.001$	-12.4%	ANOVA	$p < 0.001$	-8.6%
TPS	ANOVA	$p < 0.001$	66.1%			

The SE experienced a 62.5% increase, and the TE, a 69.4% increase in CPU utilization with the additional layer of virtualization software in place. The amount of overhead caused by virtualization can vary based on the VM's workload capable of running directly on a physical processor and the amount requiring virtualization [54]. Tudor [55] found increases in CPU utilization ranging from 38% to 45% with open source RDBMS, attributing the increase to increased I/O wait times. The current research found increases occurring in user and system CPU utilization. A comparison of bare metal and virtualized NoSQL environments found that CPU utilization increased by roughly 29% under mixed (read/write) loads [56]. In experiments using CPU-intensive benchmark applications, Pousa and Rufino [57] found decreases in CPU efficiency due to the existence of an ESXi 6.0 virtualization layer in areas of process creation, disk to RAM transfers, context switching, and system call overhead. The current research indicates that the impact of virtualization on CPU utilization on the NewSQL RDBMS was greater than levels found in open source RDBMSs and NoSQL database testing in the studies referenced. The increased CPU utilization should be strongly considered in the migration or implementation of NewSQL on virtualized systems.

The results of the ANOVA for memory utilization for the SM node indicated that the difference in the means was statistically significant, $F(1,40)=25.746$, $p<0.001$, leading to the acceptance of the alternative hypothesis that there was a difference in memory use when virtualization is implemented. A comparison of the mean values of memory utilization between bare metal (M=4702.894, SD=56.362) and virtualized (M=4766.631, SD=11.696) environments indicated only a

slight increase of 1.4% in the SM node. In the case of the TE node, the assumptions necessary to use ANOVA were not met, and a Kruskal Wallis H test was used. The results of the Kruskal Wallis H indicated that the difference in the means was statistically significant, $\chi^2(1)=23.694$, $p<0.001$, and the null hypothesis could be rejected. A comparison of the mean values of memory utilization between bare metal (M=3163.598, SD=37.606) and virtualized (M=3249.873, SD=38.754) environments indicated an increase of 2.7% in the TE node.

Concerning memory utilization, the difference between the two environments, bare metal and virtualized, indicated an increase in memory utilization in the virtualized environment of 1.4% for the SM and 2.7% for the TE. The additional overhead is present and can be attributed to the addition of the virtualization layer, but given the low percentage increases, adjustments in memory size are not warranted.

In the analysis of disk I/O, the assumptions to use an ANOVA were not met in either the case of the SM or the TE, so a Kruskal Wallis H test was used. The results of the Kruskal Wallis H indicated that the means for the SM node were significantly different, $\chi^2(1)=30.767$, $p<0.001$, and the null hypothesis was rejected. A comparison of the mean values of disk I/O utilization between bare metal (M=3574.681, SD=358.112) and virtualized (M=4907.644, SD=60.220) environments for the SM indicated an increase of 37.3%. In the case of the TE node, the Kruskal Wallis H test yielded $\chi^2(1)=1.339$, $p=0.247$. The data failed to provide the evidence needed to reject the null hypothesis, and it was therefore retained as $p>0.05$.

A statistically significant difference in the disk I/O measurement was reflected in increased disk I/O in the virtualized SM node. Using system benchmark performance tools, Pousa and Rufino [57] found disk performance under ESXi virtualized conditions to be very similar to bare metal. Shirinbab, Lundberg [56] found disk I/O write rates to be the same or higher with virtualized NoSQL databases. Tudor [55] found that disk I/O was higher in virtualized open source RDBMS through the collection of values of OS disk I/O as percentages. The faster data can be moved from disk to RAM, the greater availability of data for the RDBMS. The disk-intensive nature of RDBMSs depends on I/O bandwidth to function properly [58]. Lee and Fox [59] suggested that greater IOPS are good for database systems. The existence of statistical significance in datasets collected on the SM node allowed for the comparison of bare metal and virtualized NewSQL environments and a 37.3% increase in the disk I/O rate was observed in the virtualized environment. The increase in disk I/O can be attributed to the VMware Paravirtualized SCSI controller as noted by Dakic [51], although the increase in disk I/O in this research exceeded the 12% found in that research effort which used vSphere 6.0. Additional statistical analysis on data collected as part of the current research effort found that along with increased disk I/O rate as measured in KB/s, the virtualized SM node had increased I/O operations per second (IOPS). The virtualized SM IOPS (M=175.78, SD=2.12) exceeded bare metal SM IOPS (M=132.36, SD=10.88) by 32.8%, and the difference in the means of the two datasets was shown to be statistically significant via a

Kruskal Wallis H test, $\chi^2(1)=30.767$, $p<0.001$. The data indicates an increase in disk I/O rate under virtualized conditions for the SM node, whose role is to manage and maintain complete, consistent, independent copies of the entire database. In a discussion with a VMware engineer, he stated that the improvements emphasize the importance of disk I/O drivers in the software and that VMware drivers will coalesce disk I/O reads and writes, but the proprietary nature of the software prevented extensive discussion (D. Robertson, personal communication, May 6, 2020). Given disk I/O is often a bottleneck for increased system performance, the increases found is positive. The disk I/O data collected on the TE would not allow the rejection of the null hypothesis that stated there was no effect due to virtualization. It is worth noting that given that the role of the TE is to cache database tables in memory, accept database requests and execute SQL queries, the values of disk I/O KB/s recorded were in the single digits, which would decrease the potential impact on the system overall.

The results of the ANOVA performed on data collected for network I/O measurements indicated the difference in the means were statistically significant both for the SM, $F(1,40)=238.429$, $p<0.001$, and the TE, $F(1,40)=113.81$, $p<0.001$, and the null hypothesis was rejected for both nodes. A comparison of the mean values of network I/O between bare metal ($M=14270.151$, $SD=396.245$) and virtualized ($M=12502.987$, $SD=343.571$) environments indicated a decrease of 12.4% for the SM node. The TE node saw a difference between bare metal ($M=14491.517$, $SD=404.175$) and virtualized ($M=13243.229$, $SD=352.401$) environments corresponding with a decrease of 8.6%.

Both the virtualization of CPU resources and virtualized network adaptors will increase the time to transmit data packets [54]. The components required in the processing of virtualized network I/O are virtual network drivers, known as the vNIC, the vSwitch, the VMkernel, and the physical NIC driver [60]. This results in virtualization overhead impacting three of the four networking-specific components involved. The benchmarking server remained in a bare metal state for all experiments to ensure that virtualization effects were confined to the database nodes. The SM and the TE experienced a decrease in the network I/O (KB/s) of 12.4% and 8.6%, respectively. The statistical analysis of the data collected allowed for the acceptance of the alternate hypothesis that the additional overhead imposed is due to the virtualization of the NewSQL RDBMS. Direct-path I/O was not supported by the NIC used in the systems tested, but if available, would provide a means to minimize these performance decreases.

The fifth and final hypothesis sought to provide focus on the transactional volume of the NewSQL database. Since the necessary assumptions were met, a One-Way ANOVA was performed on the datasets using SPSS. The results of the ANOVA indicated that there was a statistically significant difference in the means, $F(1,40) = 683.821$, $p < 0.001$, indicating the null hypothesis could be rejected and the alternative hypothesis accepted. A comparison of the mean values of TPS between bare metal ($M=16.474$, $SD=1.189$) and virtualized ($M=27.821$, $SD=1.534$) environments indicated an increase of 66.1%.

A relevant, overall throughput metric for an RDBMS is the rate of database transactions, TPS. The TPS values, as reported by the benchmarking software, increased 66.1% in the virtualized environment as compared to bare metal. The improvement in TPS found in this research contradicts the results of Tudor [55], but it must be pointed out that the virtualization software was VMware ESXi 5.0 as compared to version 6.5 used in the current research. Essential to the performance of an RDBMS is the ability to quickly and efficiently read data from disk into memory when needed and to write new or updated data from memory to disk [61]. The 37.3% increase in the disk I/O and the ample CPU cycles present on the SM node provided an environment with increased processing potential. Such conditions could give the system an increased ability for transaction throughput, but additional research will be required for this to be definitively demonstrated.

V. CONCLUSION

Virtualization continues to play an important role in the delivery of IT Services in on-premise data centers and via the cloud [4]. With the virtualizing of computing resources, organizations have been able to enhance and improve the management and utilization of IT resources. New architectural approaches found in NewSQL allow for the use of SQL and adherence to relational database standards, characteristics, and guarantees in the context of the immense volumes of data present in Big Data applications [2]. Just as the use of traditional relational databases and NoSQL technologies in virtualized environments occurred, the use of NewSQL in a virtualized environment is to be expected [25]. The absence of literature reflecting research to quantify the performance impact of virtualization on NewSQL RDBMS served as a motivation for this effort. The work presented allows for a better understanding and quantification of this impact of virtualization, providing benefits to organizations seeking to virtualize NewSQL servers, and represents an effort to fill the gap in the literature on this specific topic.

The evidence of the virtualization penalty in RDBMSs [55] and NoSQL database systems [56] is present in the literature. However, with the advent of NewSQL technology and continuous improvement in virtualization software, existing paradigms should be revisited. In this research, non-trivial virtualization penalties were identified in CPU utilization and network I/O, but memory utilization was only nominally impacted, and both disk I/O and TPS values were improved in the virtualized environment. The performance improvements would indicate that the new architecture NewSQL solutions may involve dynamics different than those present in traditional and NoSQL database solutions. The architecture of the Nuodb NewSQL RDBMS creates a dependency on disk I/O for the SM, but not the TE, which is memory-dependent. Given the virtualization goal of sharing underutilized hardware resources between virtual machines, existing paradigms must be reconsidered in light of ideas such as the complementary nature of the needs of the SM and TE, which might be amendable to separate VMs on the same physical machine. Additional research surrounding such synergies should be explored.

REFERENCES

- [1] O. Ylijoki and J. Porras, "Perspectives to definition of Big Data: A mapping study and discussion," *Journal of Innovation Management*, vol. 4, no. 1, pp. 69-91, 2016, doi: 10.24840/2183-0606_004.001_0006.
- [2] A. Pavlo and M. Aslett, "What's really new with NewSQL?," *ACM SIGMOD Record*, vol. 45, no. 2, pp. 45-55, 2016, doi: 10.1145/3003665.3003674.
- [3] A. Dhanapal, M. V. Saravanakuma, and M. Sabibullah, "Emerging Big Data storage architectures: A new paradigm," *i-manager's Journal on Pattern Recognition*, vol. 4, no. 2, pp. 31-41, 2017, doi: 10.26634/jpr.4.2.13732.
- [4] H. Ur Rahman, G. Wang, J. Chen, and H. Jiang, "Performance evaluation of hypervisors and the effect of virtual cpu on performance," in *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI)*, Guangzhou, China, 2018: IEEE, pp. 772-779, doi: 10.1109/SmartWorld.2018.00146.
- [5] I. Pogarcic, D. Krnjak, and D. Ozanic, "Business benefits from the virtualization of an ICT infrastructure," *International Journal of Engineering Business Management*, vol. 4, no. Godište 2012, pp. 4-42, 2012.
- [6] I. Odun-Ayo, O. Ajayi, and C. Okereke, "Virtualization in Cloud Computing: Developments and trends," in *2017 International Conference on Next Generation Computing and Information Systems (ICNGCIS)*, Jammu, India 2017: IEEE, pp. 24-28, doi: 10.1109/ICNGCIS.2017.10.
- [7] M. Wardat, M. Al-Ayyoub, Y. Jararweh, and A. A. Khreishah, "Cloud data centers revenue maximization using server consolidation: Modeling and evaluation," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018: IEEE.
- [8] R. Kumar and S. Charu, "An importance of using virtualization technology in cloud computing," *Global Journal of Computers & Technology*, vol. 1, no. 2, pp. 56-60, 2015.
- [9] H. Fatima and K. Wasnik, "Comparison of SQL, NoSQL and NewSQL databases for Internet of Things," in *2016 IEEE Bombay Section Symposium (IBSS)*, Baramati, India, 2016: IEEE, pp. 1-6, doi: 10.1109/IBSS.2016.7940198.
- [10] S. Binani, A. Gutti, and S. Upadhyay, "SQL vs. NoSQL vs. NewSQL-a comparative study," *Communications on Applied Electronics*, vol. 6, no. 1, pp. 1-4, 2016.
- [11] J. Oliveira and J. Bernardino, "Newsql databases: Memsql and VoltDB experimental evaluation," presented at the 9th International Conference on Knowledge Engineering and Ontology Development, Funchal, Portugal, 2017.
- [12] K. Kaur and M. Sachdeva, "Performance evaluation of NewSQL databases," in *2017 International Conference on Inventive Systems and Control (ICISC)* Coimbatore, India 2017: IEEE, pp. 1-5, doi: 10.1109/ICISC.2017.8068585.
- [13] M. Murazzo, P. Gómez, N. Rodríguez, and D. Medel, "Database NewSQL Performance Evaluation for Big Data in the Public Cloud," in *Conference on Cloud Computing and Big Data*, 2019: Springer, pp. 110-121.
- [14] D. Bitton et al., "A measure of transaction processing power," *Datamation*, vol. 31, no. 7, pp. 112-118, 1985.
- [15] B. Scalzo, *Database benchmarking and stress testing: An evidence-based approach to decisions on architecture and technology*. New York, NY: Apress, 2018.
- [16] D. Rajesh, A. A. Ahmed, M. I. T. Hussan, and V. Bollapalli, "Virtualization and its role in Cloud Computing environment," *International Journal of Computer Sciences and Engineering*, vol. 7, no. 4, pp. 1131-1136, 2019, doi: 10.26438/ijcse/v7i4.11311136.
- [17] Z. Li, M. Kihl, Q. Lu, and J. A. Andersson, "Performance overhead comparison between hypervisor and container based virtualization," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*, Taipei, Taiwan, 2017: IEEE, pp. 955-962, doi: 10.1109/AINA.2017.79.
- [18] H. Fayyad-Kazan, L. Perneel, and M. Timmerman, "Benchmarking the performance of Microsoft Hyper-v server, Vmware ESXi and Xen hypervisors," *Journal of Emerging Trends in Computing and Information Sciences*, vol. 4, pp. 922-933, 2013.
- [19] B. Dordevic, V. Timčenko, N. Kraljević, and N. Davidović, "File system performance comparison in full hardware virtualization with ESXi and Xen hypervisors," in *2019 18th International Symposium INFOTEH-JAHORINA (INFOTEH)*, East Sarajevo, Bosnia, 2019: IEEE, pp. 1-5, doi: 10.1109/INFOTEH.2019.8717664.
- [20] W. Graniszewski and A. Arciszewski, "Performance analysis of selected hypervisors (virtual machine monitors-VMMs)," *International Journal of Electronics and Telecommunications*, vol. 62, no. 3, pp. 231-236, 2016, doi: 10.1515/eletel-2016-0031.
- [21] C. H. Kao, "Testing and evaluation methods for cloud environments: A review," in *Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government*, Turku, Finland, 2017: ACM, pp. 56-60, doi: 10.1145/3108421.3108435.
- [22] A. Vogelsgesang et al., "Get real: How benchmarks fail to represent the real world," in *Proceedings of the Workshop on Testing Database Systems*, Houston, TX, 2018: ACM, pp. 1-6, doi: 10.1145/3209950.3209952.
- [23] M. Aslett, "What we talk about when we talk about NewSQL," in *Too much information* vol. 2018, ed. https://blogs.the451group.com/information_management/2011/04/06/what-we-talk-about-when-we-talk-about-newsql/: the 451 group, 2011.
- [24] M. Stonebraker, "New SQL: An alternative to NoSQL and old SQL for new OLTP apps," in *BLOG@CACM* vol. 2018, ed. <https://cacm.acm.org/blogs/blog-cacm/109710-new-sql-an-alternative-to-nosql-and-old-sql-for-new-oltp-apps/fulltext>: ACM, 2011.
- [25] K. Grolinger, W. A. Higashino, A. Tiwari, and M. A. Capretz, "Data management in cloud environments: NoSQL and NewSQL data stores," *Journal of Cloud Computing: Advances, Systems and Applications*, vol. 2, no. 1, p. 22, 2013, doi: 10.1186/2192-113X-2-22.
- [26] E. Awasthi, S. Agrawal, and R. Pandey, "A survey on NoSQL and NewSQL data stores for Big Data management," *Global Journal of Engineering Science and Research*, vol. 5, no. 7, pp. 269-277, 2018, doi: 10.5281/zenodo.1313633.
- [27] A. Almassabi, O. Bawazeer, and S. Adam, "Top NewSQL databases and features classification," *International Journal of Database Management Systems (IJDMS)*, vol. 10, no. 2, pp. 11-31, 2018, doi: 10.5121/ijdms.2018.10202.
- [28] J. M. Monteiro, A. Brayner, and J. A. Tavares, "What comes after NoSQL? NewSQL: A new era of challenges in DBMS scalable data processing," in *Tópicos em Gerenciamento de Dados e Informações. Minicursos do XXXI Simpósio Brasileiro de Banco de Dados (SBBDD)*, Salvador, Brazil, 2016, pp. 27-56.
- [29] I. Hrubaru and M. Fotache, "On the performance of three in-memory data systems for on line analytical processing," *Informatica Economica*, vol. 21, no. 1, pp. 5-15, 2017.
- [30] O. Borisenko and D. Badalyan, "Evaluation of SQL benchmark for distributed in-memory database management systems," *International Journal of Computer Science and Network Security* vol. 18, no. 10, pp. 59-63, 2018.
- [31] Apache.org. "In-memory computing platform." <https://ignite.apache.org/> (accessed).
- [32] "What is VoltDB's database product?" <https://www.voltDB.com/product/> (accessed 8/5/19, 2019).
- [33] CentOS. (2019). The CentOS Project, www.centos.org. Accessed: 11/5/2019. [Online]. Available: <https://www.centos.org/>.
- [34] "What Is NuoDB?" <https://www.nuodb.com/product-overview> (accessed).
- [35] R. Verma, "Understanding the technological trends and quantitative analysis of NewSQL databases," Masters, Department of Computer Science and Electrical Engineering, University of Maryland, ProQuest Dissertations & Theses Global, 2017. [Online]. Available: <https://proxy.cecylibrary.com/login?url=https://search-proquest-com.proxy.cecylibrary.com/docview/1940284796?accountid=144789>.
- [36] NuoDB, "NuoDB Architecture." [Online]. Available: <http://go.nuodb.com/rs/099-DVI-451/images/Technical->

- Whitepaper.pdf?aliId=eyJpJoiME9hNmV4OXZyWkRUYzZZbSIsInQiOiJWbWlDMVFQK0V4VFBGalJBM1g2THJ3PT0ifQ%253D%253D.
- [37] NuoDB. "Installing NuoDB on Linux." <http://doc.nuodb.com/Latest/Content/Nuoadmin-Installing-NuoDB-on-Linux.htm> (accessed 12/15/19).
- [38] nmon for Linux. (2017). [Online]. Available: <https://sourceforge.net/projects/nmon/>.
- [39] M. F. Khalid, B. I. Ismail, and M. N. M. Mydin, "Performance comparison of image and workload management of edge computing using different virtualization technologies," *Advanced Science Letters*, vol. 23, no. 6, pp. 5064-5068, 2017, doi: 10.1166/asl.2017.7310.
- [40] M. Kaczmarek, P. Perry, J. Murphy, and A. O. Portillo-Dominguez, "In-test adaptation of workload in enterprise application performance testing," in *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering Companion*, 2017: ACM, pp. 69-72.
- [41] D. E. Difallah, A. Pavlo, C. Curino, and P. Cudre-Mauroux, "Otpbench: An extensible testbed for benchmarking relational databases," *Proceedings of the VLDB Endowment*, vol. 7, no. 4, pp. 277-288, 2013.
- [42] J. Darmont, "Data processing benchmarks," *Encyclopedia of Information Science and Technology*, pp. 1741-1747, 2014, doi: 10.4018/978-1-4666-5888-2.ch167.
- [43] F. Coelho, J. Paulo, R. Vilaça, J. Pereira, and R. Oliveira, "Htapbench: Hybrid transactional and analytical processing benchmark," in *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering*, 2017: ACM, pp. 293-304.
- [44] I. Kovacevic and I. Mekterović, "Alternative business intelligence engines," in *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 2017: IEEE, pp. 1385-1390.
- [45] D. Seybold and J. Domaschka, "Is Distributed Database Evaluation Cloud-Ready?," in *European Conference on Advances in Databases and Information Systems*, 2017: Springer, pp. 100-108.
- [46] R. Cole et al., "The mixed workload CH-benCHmark," in *Proceedings of the Fourth International Workshop on Testing Database Systems*, 2011: ACM, p. 8.
- [47] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences.," *Behavior Research Methods*, vol. 39, pp. 171-191, 2007.
- [48] M. Raasveldt, P. Holanda, T. Gubner, and H. Mühleisen, "Fair benchmarking considered difficult: Common pitfalls in database performance testing," in *Proceedings of the Workshop on Testing Database Systems*, 2018: ACM, p. 2.
- [49] VMware.com. "VMware.com." VMware.com. <https://www.vmware.com/> (accessed 7/19/2019, 2019).
- [50] G. Chen, "Worldwide Virtual Machine Software MarketShares, 2017: Virtualization Still Showing Positive Growth," *International Data Corporation*, 2018. [Online]. Available: <https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/products/vmware-idx-virtual-machine-market-shares-2017.pdf>
- [51] V. Dakic, "Influence of virtual storage controller type on Microsoft SQL Server 2019 performance," *Annals of DAAAM & Proceedings*, vol. 30, 2019.
- [52] B. Goldsand and C. Brown, "SAP® Sybase® Adaptive Server Enterprise on VMware vSphere® essential deployment tips," 2013.
- [53] VMware, "vSphere Virtual Machine Administration." [Online]. Available: <https://docs.vmware.com/en/VMware-vSphere/6.5/vsphere-esxi-vcenter-server-65-virtual-machine-admin-guide.pdf>
- [54] VMware, *Performance Best Practices for VMware vSphere 6.5*, Palo Alto: VMware, 2017. [Online]. Available: https://www.vmware.com/content/dam/digitalmarketing/vmware/en/pdf/techpaper/performance/Perf_Best_Practices_vSphere65.pdf.
- [55] A. Tudor, "A study of open-source relational database performance in virtual computing environments," *Doctoral Dissertation*, Northcentral University, Ann Arbor, 2014. [Online]. Available: <https://login.ctu.idm.oclc.org/?url=http://search.proquest.com/docview/1506579000?accountid=26967>.
- [56] S. Shirinbab, L. Lundberg, and E. Casalicchio, "Performance evaluation of container and virtual machine running Cassandra workload," in *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*, Rabat, Morocco 2017: IEEE, pp. 1-8, doi: 10.1109/CloudTech.2017.8284700.
- [57] D. Pousa and J. Rufino, "Evaluation of type-1 hypervisors on desktop-class virtualization hosts," *Iadis Journal on Computer Science and Information Systems*, vol. 12, no. 2, pp. 86-101, 2017.
- [58] M. G. Xavier, K. J. Matteussi, F. Lorenzo, and C. A. De Rose, "Understanding performance interference in multi-tenant cloud databases and web applications," in *2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, 2016: IEEE, pp. 2847-2852, doi: 10.1109/BigData.2016.7840933.
- [59] H. Lee and G. Fox, "Big Data Benchmarks of High-Performance Storage Systems on Commercial Bare Metal Clouds," in *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, 2019: IEEE, pp. 1-8.
- [60] D. B. Oljira, A. Brunstrom, J. Taheri, and K.-J. Grinnemo, "Analysis of network latency in virtualized environments," in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016: IEEE, pp. 1-6.
- [61] R. Elmasri and S. Navathe, *Fundamentals of database systems*, 7th ed. Boston, MA: Pearson, 2017.

Knock Knock, Who's There: Facial Recognition using CNN-based Classifiers

Qiyu Sun, Alexander Redei
Department of Computer Science
Central Michigan University
Mount Pleasant, MI

Abstract—Artificial intelligence (AI) has captured the public's imagination. Performance gains in computing hardware, and the ubiquity of data have enabled new innovations in the field. In 2014, Facebook's DeepFace AI took the facial recognition industry by storm with its splendid performance on image recognition. While newer models exist, DeepFace was the first to achieve near-human level performance. To better understand how this breakthrough performance was achieved, we developed our own facial image detection models. In this paper, we developed and evaluated six Convolutional Neural Net (CNN) models inspired by the DeepFace architecture to explore facial feature identification. This research made use of the You Tube Faces (YTF) dataset which included 621,126 images consisting of 1,595 identities. Three models leveraged pretrained layers from VGG16 and InceptionResNetV2, whereas the other three did not. Our best model achieved a 84.6% accuracy on the test dataset.

Keywords—Face recognition; deep learning; convolutional neural networks; DeepFace

I. INTRODUCTION

Facial recognition is a method of identifying an individual using his or her face from a digital image or a video clip. Such methods could be used for facial authentication by pinpointing and determining facial features from a given image, uniquely identifying the person. Initially this was limited to desktop computers due to demanding computational power constraints. Recently however it has seen wider usage, such as on mobile devices, robotics, finding missing people, and diagnosing diseases. Facial recognition is also applied in diagnosing diseases. 22q11.2 deletion syndrome (22q11.2 DS) is the most common micro-deletion syndrome and was underdiagnosed in a variety of populations in the past. Because the disease results in multiple defects throughout the body, including cleft palate, heart defects, a characteristic facial appearance, and learning problems, healthcare providers often can't pinpoint the disease, especially in diverse populations. After analyzing the disease with facial analysis technology, researchers found that sensitivity and specificity were greater than 96% for all populations, which demonstrated how facial analysis technology can assist clinicians in making accurate 22q11.2 DS diagnoses [1]. Researchers with the National Human Genome Research Institute (NHGRI), part of the National Institutes of Health, and their collaborators, have successfully used facial recognition software to diagnose a rare, genetic disease in Africans, Asians, and Latin Americans [2].

Facial recognition technology has a wide range of applications and profound social and cultural impacts and has been introduced across various aspects of public life. For

example, facial recognition payment services are now possible. Nowadays, in China people can purchase food at the grocery store and can even complete their payment directly by scanning their face at a register without needing a credit card or mobile application [3]. In terms of the design and implementation of security systems, facial recognition technology also has a wide range of applications including web and mobile authentication [4], airport check-in [5], and smart medicine cabinets [6]. In the education industry, facial recognition has been applied to compulsory schooling to address issues such as campus security, automated registration, and student emotion detection and has largely been seen as routine additions to school systems with already extensive cultures of monitoring and surveillance [7]. While facially driven learning has been widely used, critical commentators are beginning to question the pedagogical limitations of it. They purposed multiple questions about facial recognition technology including the likelihood of it altering the nature of schools and schooling along divisive, authoritarian and oppressive lines, and what kind of law and regulatory mechanisms can help for eliminating the potential risks to consumers when they are making use of it [7]. Due to the relatively limited technology, the current ability to detect human faces in this field provides a buffer from coping with the potential consequences including a serious threat to online identities being misused by hackers for illegal activities.

An overview of the rest of the paper is as follows: in Section 2 we reviewed some of the related work in the same research field with DeepFace; in Section 3, we introduce core techniques related to DeepFace: Deep Learning and Convolutional Neural Networks; Section 4 describes the 3D model-based face alignment method applied and the model architecture used; Section 5 talks about our deep learning model that follows the architecture of DeepFace's and was trained on YouTube Faces (YTF) video data set. In Section 6 we present some quantitative results of our models and the last section is the conclusion.

II. RELATED WORK

Biometric facial recognition, also known as automatic face recognition, is a particularly attractive method of biometric recognition because it focuses on "faces," the same identifiers that humans primarily use to distinguish people. One of its main goals is the understanding of the complex human visual system and the knowledge of how humans represent faces in order to discriminate different identities with high accuracy. Facial recognition consists of three basic processes: detection, capture, and face match. The detection process is to determine

if there is a target face in the source. The capture process transforms the targeted face into a set of digital data based on the facial features. The face match process verifies if the two faces are of the same person. This process is shown in Fig. 1 below.

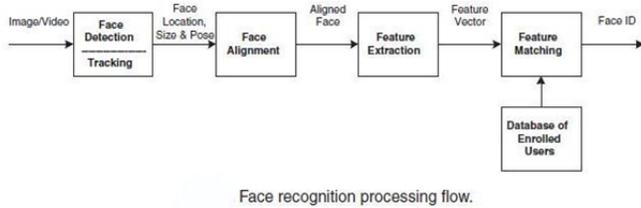


Fig. 1. Facial Recognition Processing Flow.

A large number of approaches have emerged in the field of facial recognition, including a hand-crafted features based method [8] and a widely applied metric learning methods with task-specific objectives [9]. These approaches were never quite able to reach human-level performance in identifying faces. Although progress in facial recognition was encouraging, the task has also turned out to be a difficult endeavor.

A. DeepFace

DeepFace is a deep learning face recognition technology developed by a research group at Facebook. It identifies human faces in digital images with human-level performance. In DeepFace, researcher revisited both the alignment step and the representation step of the face recognition process and proposed a new approach of deriving a face representation by employing explicit 3D face modeling. It employed a nine-layer neural net with over 120 million connection weights and was trained on four million images uploaded by Facebook users [10] [11]. DeepFace demonstrates that a 3D model-based alignment method can effectively help in face recognition and closes the gap to human-level accuracy. Next Generation Identification (NGI) is another application developed by Federal Bureau of Intelligence (FBI) of the same year. According to one report the NGI's performance is non-satisfactory. It returns a ranked list of 50 possibilities and only promises an 85% chance of returning the suspect's name in the list [12]. The DeepFace system (stated by the Facebook Research team) reaches an accuracy of $97.35 \pm 0.25\%$ on labeled faces in the wild (LFW) data set whereas human beings have 97.53% [13]. Google FaceNet later achieved a 99.65% accuracy on the same data set [14].

B. Local Binary Patterns (LBP)

Local binary patterns (LBP) is a type of visual descriptor used for classification in computer vision and it is the particular case of the Texture Spectrum model proposed in 1990 [15]. LBP was first described in 1994 [16] and it is a simple yet very efficient texture operator which labels the pixels of an image by thresholding the neighborhood of each pixel and considers the result as a binary number [17]. Using the LBP combined with histograms we can represent the face images with a simple data vector [18]. In the LBP approach for texture classification, the occurrences of the LBP codes in an image are collected into a

histogram. The classification is then performed by computing simple histogram similarities. However, considering a similar approach for facial image representation results in a loss of spatial information and therefore one should codify the texture information while retaining also their locations. One way to achieve this goal is to use the LBP texture descriptors to build several local descriptions of the face and combine them into a global description. The basic methodology for LBP based face description proposed by Ahonen et al. [19] is as follows: The facial image is divided into local regions and LBP texture descriptors are extracted from each region independently. The descriptors are then concatenated to form a global description of the face, as shown in Fig. 2.

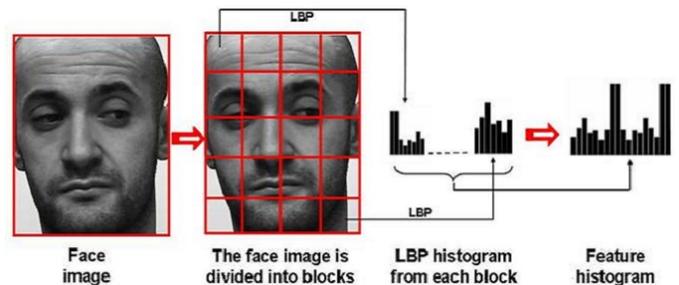


Fig. 2. Face Description with Local Binary Patterns.

C. DeepID-Net

A pre-trained model is a model that was trained on a large benchmark dataset to solve a problem similar to the one that we want to solve. As for most image detection problems, the main features of the objects to be detected are often similar, so a pre-trained model can be leveraged to typically get improved performance. But researchers found a gap between the pre-training task and the fine-tuning task that makes pre-training less effective [20]. Inspired by the need to adopt more targeted optimization solutions for specific objects, researchers propose the DeepID-Net model. DeepID-Net is an image detection model developed by the Multimedia Laboratory of the Chinese University of Hong Kong. Its full name is DeepID-Net: Deformable Deep Convolutional Neural Networks for Object Detection [20]. They added more steps on the region-based convolutional neural networks on the region-based convolutional neural networks (R-CNN) processing, including bounding box rejection, deep model training, def pooling layer, SVM-net(replace softmax with hinge loss to accelerate learning), multi-stage training, etc, as shown in Fig. 3. The model yields a 99.8% accuracy, while the state-of-the-art method achieves a 97% accuracy when testing multi-view facial images [20]. This paper was published on CVPR2014. After that, the team focused on applying the model to the specific application of facial recognition, and correspondingly made some changes and optimizations to the DeepID model. The updated two versions of the model are called DeepID2 and DeepID3.

D. FaceNet

FaceNet is a universal system that can be used for face verification (is it the same person?), recognition (who is this person?) and clustering (looking for similar people?) [22]. The

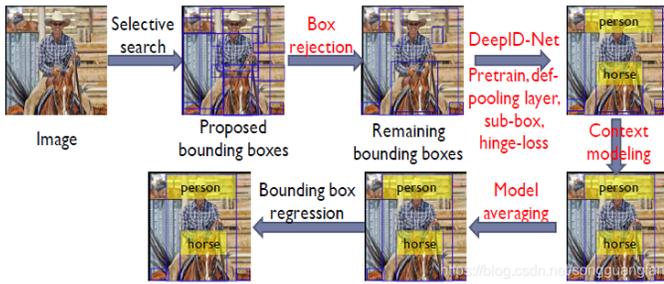


Fig. 3. Overview of DeepID-Net Process (Texts in Red Highlight the Steps that are Not Present in RCNN.) [21]

method adopted by FaceNet is to map images into Euclidean space through a convolutional neural network. Different from the application of other deep learning methods on human faces, FaceNet did not use the traditional softmax method to classify and learn, and then extract a certain layer as a feature. It directly used triplets-based LMNN (Maximum Boundary Neighbor Classification) loss function to train the neural network, and the network directly outputs a 128-dimensional vector space [22]. FaceNet has achieved an accuracy of $99.63 \pm 0.09\%$ on the LFW dataset and an accuracy of $95.12 \pm 0.39\%$ on the YTF dataset [22]. The advantage of this model is that the target image can be used with very little processing. It also provides future research directions, such as analyzing wrong samples to improve accuracy, reducing model size to speed up training, etc.

III. METHODOLOGY

Deep learning is a specific subfield of machine learning [23]. It represents learning process from data, emphasizing on learning successive "layers" of increasingly meaningful representations. The word "deep" in "deep learning" is not referring to deeper understanding achieved through the approach but stands for the idea of successive layers of representations. The number of layers that contribute to the model is called the depth of the model. These layered representations are learned through models called neural networks and they are structured in layers stacked one after the other. Deep learning is technically a mathematical framework for learning representations from data with a multi-stage way. A large deep network has multiple layers with many more nodes in each layer, which leads to many more parameters to tune. It would be too slow and insufficient to train a deep learning model without a large dataset and powerful computers. Compared to the traditional learning algorithms (Regression, Random Forest, Support Vector Machine, etc.), deep learning may not necessarily outperforms when given data of small scale. But once the data scale goes up exponentially, deep learning outperforms others because more parameters provide the capability to learn complicated nonlinear patterns [24]. Generally, we expect the model to capture the most helpful features by itself without too much expert-involved manual intervening on features learning.

Machine learning is about mapping inputs to target outputs. The specification of what each layer does to their input data is stored in a bunch of parameters called "weight". The learning process refers to finding a set of values of the weights of all layers in a network so that the network will correctly map

inputs to their associated targets. But here comes the issue: to find the correct value for all of the weights can be a daunting task, especially when modifying the value of one parameter will affect the performance of the whole model. To control the output of a neural network, the loss function plays an important role in making the prediction of the network and the target. It computes a distance score measuring how well the network performs. The job of the "optimizer" is to use this score as a feedback signal to adjust the value of the weights, successively trying to lower the loss score. Implementing "back-propagation" is the central algorithm used for this in deep learning architectures. In recent years, deep learning has achieved a revolution with tremendous achievements on many types of difficult problems, especially perceptual problems, which have long been historically difficult for machine learning.

Convolutional neural networks are a type of feed-forward artificial neural networks, most commonly applied to analyzing visual imagery [25]. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on their shared-weights architecture and translation invariance characteristics [26] [27]. They have applications in image classification, Image segmentation, character recognition [28], medical image analysis [29], natural language processing [30].

A convolutional neural network consists of an input layer, multiple hidden layers and an output layer. In any feed-forward neural network, all middle layers are called hidden layers due to their inputs and outputs are sealed by the activation function and final convolution [31]. Convolution refers to a mathematical operation between two matrices, it is defined as the integral of the product of the two functions after one is reversed and shifted. It then evaluates the integral over all values of the shift to produce a convolutional function. Convolutional networks are a specialized type of neural networks that use convolution in place of general matrix multiplication in at least one of their layers [32]. The convolutional layer has a defined fixed small matrix, also called a kernel or filter. It computes the element-wise multiplication of the values in the kernel matrix and the original image values as the kernel is sliding, or convolving, across the matrix representation of the input image as shown in Fig. 4. Specially designed kernels can fast and efficiently process images for common purposes like edge detection and many others. Convolutional and pooling layers respond to feature extraction. The fundamental difference between a densely connected layer and a convolution layer is that dense layers learn global patterns in their input feature space, while convolution layers learn local patterns [33].

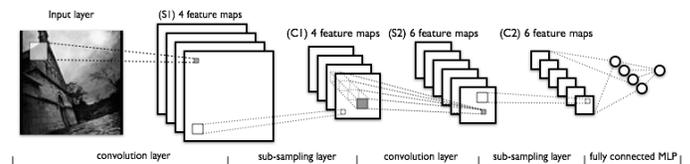


Fig. 4. The LeNet Architecture Consists of Two Sets of Convolutional, Activation, and Pooling Layers, Followed by a Fully-connected Layer, Activation, Another Fully-Connected Layer, and Finally a Softmax Classifier [34].

We used the development environment provided by Google Colab with the TensorFlow and Keras stacks. All the models

were trained using NVIDIA Tesla T4 graphical processing units (GPU) hardware equipped with 16 GB of memory and 12.7 GB RAM. This allowed us to create and simulate a deepface-esq model architecture.

For the interested reader, a link to our source code repository can be found here: <https://github.com/QiyuSun/Facial-Recognition-using-CNN-classifier>. This repository includes our Jupiter notebook python files as well as a readme file with links to the YTF dataset. The full dataset is not included in our repo due to space constraints, we only link to it.

IV. DEEPFACE ARCHITECTURE

Before the deepface architecture came about, one of the challenges to facial recognition was the reduced accuracy caused by face images collected from different perspectives. In fact, facial alignment is still considered a difficult issue, especially in an unsupervised environment. The task of face alignment is to automatically locate key facial feature points, such as eyes, nose tip, mouth corners, eyebrows, and contour points of various parts of the facial contour according to the input face image. The process of face alignment can be divided into three sub-problems: 1) How to model the apparent image (input) of a human face? 2) How to model the face shape (output)? 3) How to establish the association between the apparent image (model) of the face and the shape (model) of the face? In terms of the DeepFace method, Facebook researchers have made a great contribution in the development of an effective deep neural network architecture with a very large, labeled dataset of faces, an effective facial alignment system based on explicit 3D modeling of faces, and results that reach near real time human-level performance [13].

A. Alignment Pipeline

The alignment pipeline of DeepFace is as follows:

- (a) Detect face with 6 initial points.
- (b) Crop out the face with 2D-aligned inducing.
- (c) Apply Delaunay triangulation by 67 fiducial points on the 2D-aligned crop, adding triangles on the contour to avoid discontinuities.
- (d) Transform triangulated face into 3D shape.
- (e) The face becomes a deep 3D triangle net.
- (f) Deflect the triangulation.
- (g) The final frontalized crop.
- (h) A new view generated by the 3D model (not used in paper).

The function of these steps uses the 3D model to align the face, so that the CNN can exert its maximum effect. This is shown in Fig. 5.

B. Representation

After 3D alignment, the images formed are all shrunk into 152x152 pixel inputs into the network structure shown in Fig. 6, the parameters of the structure are as follows:

- (a) The 3D aligned 152x152 pixel 3-channel RGB face image is sent to the convolutional layer (C1) with 32x11x11x3 filters.

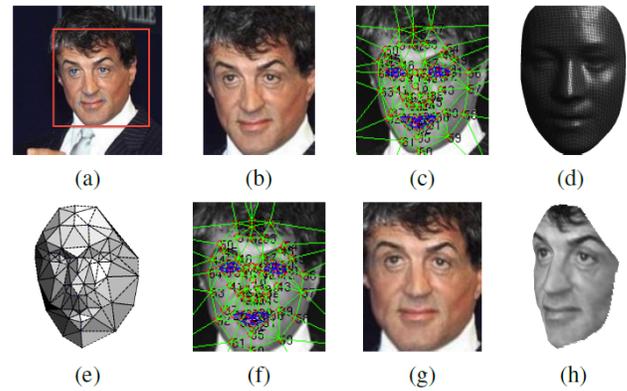


Fig. 5. DeepFace Alignment Pipeline.

- (b) Then the obtained 32 feature maps are fed to the maximum pooling layer (M2), and the 3x3 spatial neighborhood is maximum pooled with a stride of 2. Each channel is executed separately.
- (c) After M2 is a convolutional layer with 16x9x9x16 filters (C3).
- (d) Locally connected 1 [35], but each position in the feature map learns a different filter bank. Local means the parameters of the convolution kernel do not share. Locally connected layer is different from the convolutional layer in its kernel.
- (e) F7 and F8 are fully connected layers and they can capture the correlation between the features of the face image, such as the position and shape of the eyes and mouth. The output of F7 will be used as the original face representation feature vector with 4096 dimensions. The face representation on F8 is sent to K-way Softmax to generate the probability distribution on the category label for classification. The 4030 dimension is respective to the number of identities in the SFC training data set, and each identity has 800 to 1200 face pictures.
- (f) Normalization of face representation: Normalize the face representation feature to be between zero and one to reduce the sensitivity to changes in illumination.

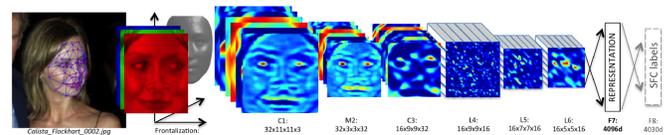


Fig. 6. Outline of the DeepFace Architecture.

C. Datasets

In DeepFace, researchers trained their model on the Social Face Classification (SFC) dataset and evaluated the model

on the Labeled Faces in the Wild database (LFW) and the YouTube Faces (YTF) dataset. In this work, our model is built on the YTF dataset. The YTF Database is a face video database, which aims to study the problem of unconstrained face recognition in videos. It contains 3,425 videos of 1,595 different identities. All the videos were downloaded from YouTube. It provides an average of 2.15 videos available for each subject with clips duration varying from 48 frames to 6,070 frames, and 181.3 frames of average length. It initially performs automatic screening to ensure that the videos are long enough to capture useful information for the various recognition algorithms with stable detection. The remaining videos were manually verified to ensure that the videos would be correctly labeled corresponding to the subjects, not static images or slides, and no duplicated videos were included [36].

In terms of designing the data set structure and benchmarks, the YTF dataset follows the principal of the Labeled Faces in the Wild (LFW) collection. All video frames are encoded with well-built descriptors with the face detector output considered in each frame. The face images are bounded and cropped from the frame, 2.2 times of their original sizes. Additionally, the images are resized to 200x200 pixels then cropped into 100x100 pixels in central area. The images are aligned by fixing the coordinates of facial feature points following a conversion to grayscale. The image is divided to a fixed grid of blocks with the descriptions of each block normalized to a unit Euclidean length [36]. For the benchmark tests of the YTF dataset, the YTF dataset follows the example of the LFW benchmark various tests like standard test and ten-fold test. It is divided into 5,000 video pairs and 10 groups, for evaluating video-level face verification [36].

V. MODEL TRAINING AND EVALUATION PROTOCOL

Six models were built and trained. First we built a baseline model, next a frame base model, aligned base model, VGG16 base model, InceptionResNetV2 base model, and finally InceptionResNetV2 model. The training and validation distribution for all models followed the same split.

A. Dataset

Twenty images were extracted for the train set of each identity. The remaining images in folders of each identity are divided into training set and validation set with ratio of 8:2. All base models were built on a subset of YTF frame_images_DB and aligned_image_DB with 160 classes, which consisted of the first 10 percent of videos ordered by name. It is of note that in practical face recognition applications today, the images processed by the model are often already aligned. For example in our dataset, each image is assigned a unique floating point number corresponding to its identity. For example Figure 7 corresponds to the unique identifier '0.614', where the '0' indicates the folder with the identity of a known actor — Aaron Eckhart — and '614' indicates the particular frame sequence in the dataset. Because pre-aligned images were already available, we did not implement an alignment subsystem. Instead after implementing a model architecture, we trained and validated that model on the aligned_image_DB dataset.



Fig. 7. '0.614' Images of Aaron Eckhart in Frame_Image_DB (left) and Aligned_Image_DB (Right).

B. Baseline Model

Our first approach initially applied a CNN-based model with a single Conv2D layer to train a baseline model on frame_image_DB subset for developing a better performing model. The baseline model consisted only one Conv2D layer with 32 nodes, input shape of (152, 152, 3) and activation function relu. The output layers consisted of 160 nodes and the pooling window sizes were 3 by 3 and 2 by 2 for Conv2D and MaxPooling2D, respectively. After converting the pooled feature map to a single column, only one Dense layer was defined and which also served as output layer using softmax for multi-class classification. Categorical_crossentropy, Adam, and accuracy were defined in compiling for loss, optimizer and metrics. After training, the accuracy of training and validation reached 99.84% and 97.64% within 20 epochs Fig. 8.

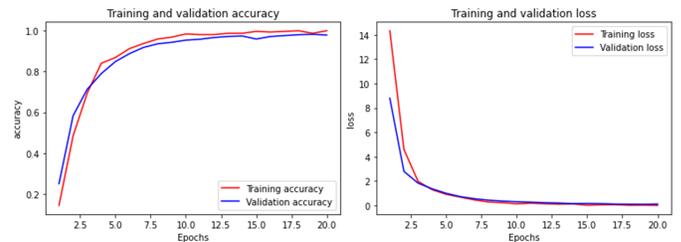


Fig. 8. Baseline Model Performance.

C. Frame and Aligned Base Model

After the baseline model was developed, we added one more Conv2D layer with 16 nodes and relu as activation function, one more dense layer with 1024 nodes to train the Frame Base Model and Aligned Base Model, essentially structuring our model to the model architecture of DeepFace. Since the data set is relatively large, and there was no obvious overfitting observed during the training process, Regularization methods were not added. The learning rate of optimizer Adam was 0.00002. Both Frame Base Model and Aligned Base Model were trained to 20 epochs. The frame base model reached 97.23% and 95.15% accuracy of training and validation Fig. 9. The Aligned Base Model reached 86.51% and 80.65% accuracy of training and validation Fig. 10.

D. VGG16 and InceptionResNetV2 Base Model

The next iteration constructed utilized the VGG16 Base Model and InceptionResNetV2 Base Model following a similar architecture. The Conv2D layers and MaxPooling layers of both two models were replaced with pre-trained conv_base.

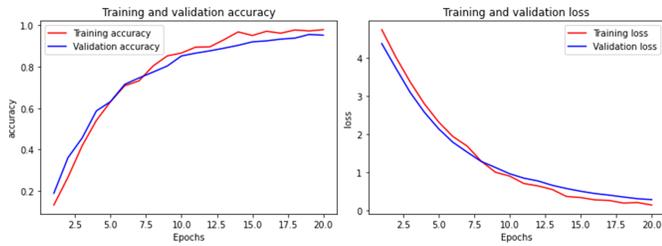


Fig. 9. Frame Base Model Performance.

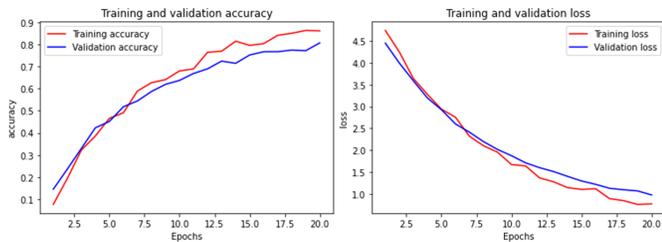


Fig. 10. Aligned Base Model Performance.

We froze the top layers of conv_base so that weights of those layers would keep unchanged during training process. The first dense layer was set 4096 nodes. After training, the VGG16 Base Model reached 98.76% and 97.03% accuracy of training and validation Fig. 11. The InceptionResNetV2 Base Model reached 99.75% and 97.23% accuracy of training and validation Fig. 12.

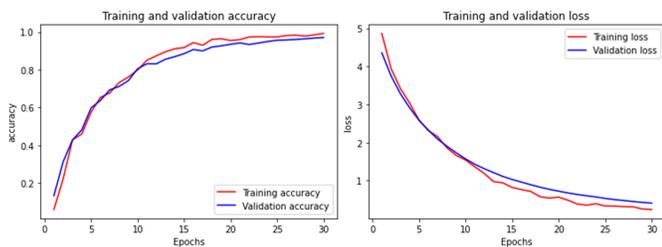


Fig. 11. VGG16 Base Model Performance.

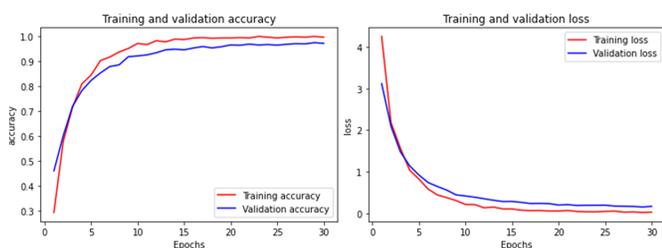


Fig. 12. InceptionResNetV2 Base Model Performance.

E. InceptionResNetV2 Model

The final InceptionResNetV2 Model shared the same model architecture with the InceptionResNetV2 Base Model. It was trained on the entire aligned_image_DB dataset (621,126 images of 1,595 identities) with same dataset distribution as

the base dataset. After training with 200 epochs, the InceptionResNetV2 model reached 91.12% and 90.79% accuracy of training and validation as shown in Fig. 13.

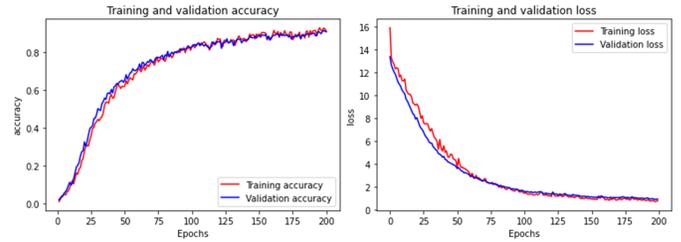


Fig. 13. InceptionResNetV2 Model Performance.

VI. RESULTS

The basic baseline model showed great performance in accuracy (99.84% on train set, 97.64% on validation set, 98.60% on test set) but it was just a simple classifier that allowed us to explore what parameters needed to be tweaked to obtain even better results. We can use the performance of other models as a baseline to evaluate the performance of the all models trained on the specific dataset. Based on the baseline models, we developed the Frame Base Model and Aligned Base Model. The Frame Base Model reached 97.23%, 95.15% and 90.78% accuracy on train set, validation set and test set. The Aligned Base Model reached 86.51%, 80.65% and 64.92% accuracy on train set, validation set and test set (Table I, Table II). The Frame Base Model considerably outperformed but the result was not that convincing. It was built on different datasets with the same architecture, which implied that the difference in performance could be caused by the specific dataset used for validation. At this point, we believe the Base Model learned little from facial features and instead was emphasizing on other factors too much. It was those no-facial related features that helped it reached greater performance than the Aligned Base Model, meaning it could be objects like chairs, studio backgrounds, etc.

The VGG16 Base Model and the InceptionResNetV2 Base Model were trained on subset of aligned image subset with more neurons added in dense layers and more training epochs compared to previous two models. The VGG16 Base Model reached 94.98% and 93.47% of train and validation accuracy at the 20th epoch, and the InceptionResNetV2 Base Model reached 99.63% and 96.58% of accuracy, respectively. After 30 epochs of training, the VGG16 Base Model reached 98.76%, 97.03% and 89.60% on train set, validation set and test set (Table I, Table II), and the InceptionResNetV2 Base Model reached 99.75%, 97.23% and 93.96% of accuracy on train set, validation set and test set (Table I, Table II). The accuracy curves of the two models were both smooth, which indicated that no apparent overfitting was observed in training. On the ground, we can see the great performance of pre-trained model layers in image classification.

The finalized InceptionResNetV2 Model was built with the same structure as the InceptionResNetV2 Base Model. It had 1,595 neurons in the final output layer corresponding to the number of identities in the completed aligned image dataset (621,126 images of 1,595 identities) and was trained

with 50 epochs. It took 0.033s for each image during the training process. The model reached accuracy of 92.75% and 91.46% at the 198th epochs then it performed increasingly higher loss and relatively lower accuracy afterwards (Table III). The decline in performance may be caused by a variety of factors, including poor architecture of model and overfitting. Considering that no explicit overfitting was found in previous models, it would be of help to promote the model performance with a better networks architecture rather than with additional regularization methods added. In the DeepFace architecture, researchers added three locally-connected layers after 3D convolutional layers and maxpooling layers, which might be one of the solutions for structural optimization of the model. In the final evaluation on test set, the InceptionResNetV2 Model performed 84.60% of accuracy and 1.2582 of loss, which demonstrated that more tuning on the model were required, as well as some pre-operations on images before being fed into model.

TABLE I. PERFORMANCE OF TRAINING & VALIDATION ACCURACY

Model	Train-Acc	Val-Acc
Baseline Model	99.84%	97.64%
Frame Base Model	97.23%	95.15%
Aligned Base Model	86.51%	80.65%
VGG16 Base Model	98.76%	97.63%
InceptionResNetV2 Base Model	99.75%	97.23%
InceptionResNetV2	91.12%	90.79%

TABLE II. PERFORMANCE ON TEST DATA

Model	Test-Loss	Test-Accuracy
Baseline Model	0.0866	98.60%
Frame Base Model	0.6357	90.78%
Aligned Base Model	1.8023	64.92%
VGG16 Base Model	0.9026	89.60%
InceptionResNetV2 Base Model	0.2992	93.96%
InceptionResNetV2	1.258	84.60%

TABLE III. INCEPTIONRESNETV2 MODEL PERFORMANCE IN FINAL EPOCHS

Epoch	Loss	Accuracy	Val_loss	Val_Accuracy
195	0.7641	0.9294	0.9466	0.9062
196	0.7965	0.9156	0.9768	0.8975
197	0.8055	0.9131	0.9380	0.9082
198	0.7046	0.9275	0.9005	0.9146
199	0.7391	0.9231	0.9204	0.9106
200	0.7834	0.9112	0.9276	0.9079

We were not able to best DeepFace's 96% accuracy. However, our top model achieved a respectable 84.6%. Considering the constrained resources we had (this was developed entirely on a single laptop and Google Colab compared to the massive resources available at Facebook), the mission of this project was achieved.

We directly used the aligned dataset published in YouTube Face dataset rather than implementing a specific face alignment method. In DeepFace, the method developed to map 2D human facial features to 3D models and use them as 3D input to train models was key to making DeepFace achieve its breakthrough outstanding performance. In addition, unlike our training and verification based entirely on the YouTube Face (YTF) dataset, DeepFace's training set and verification set involved a total of three different data sets (Social Face Classification dataset,

Labeled Faces in the Wild dataset, and YTF dataset). Training on one dataset and using the different datasets for validation reduces its accuracy deviation when facing images of different sizes and types. When looking only at the YTF dataset making full use of these factors, DeepFace achieved the test accuracy of $91.4 \pm 1.1\%$. This number is a more accurate threshold to compare our model against as it's an apples-to-apples comparison leveraging the same dataset. Taking into account the limitations of so many conditions mentioned above, the result we obtained, when compared to DeepFace, seems to be rewarding.

Although we were inspired by the architecture of DeepFace, as described in detail above we did not fully copy the DeepFace model. We were limited by the computational power available to us. Just the memory required to fully reproduce the DeepFace model is massive and greatly exceeds that which we had access to. Instead, we demonstrated that the simpler and more accessible models we built have promise in recreating DeepFace-style breakthrough performance, utilizing a fraction of the resources.

VII. CONCLUSION

DeepFace revolutionized the facial image recognition industry. In this paper, we demonstrated the power of learned features through six convolutional neural networks (CNNs). Inspired by the DeepFace architecture, but in making our own tweaks, the models we constructed were trained on the YouTube Faces (YTF) dataset to be multi-class classifiers. The Base Models showed satisfactory performance, which indicated that a CNN-based architecture could manifest remarkable performance in image classification if given a large dataset. Compared to the DeepFace model architecture, the model was less complex, which potentially would bring difficulties in capturing essential facial features effectively in other datasets. Based on the first experiment, it was obvious that there are multiple factors that needed to be taken into consideration when constructing an image classifier for a face recognition system. When dealing with exponentially massive amounts of data, the architecture and depth of the model will play a crucial role in performance. The success of DeepFace showed that remarkable results could be achieved with the right architecture combined with face alignment and frontalization. We demonstrated that our models could obtain good results at much lower computational cost.

The first few models we built showed us many factors that need to be considered when building a large CNN classifier, such as: how to make full use of the structure and characteristics of CNN itself, the suitable combination of hyperparameters for training, and how to adjust particular parts of model architecture when working with datasets at large scales. This paper is a valuable contribution to the field of image classification and facial recognition.

In future work, we wish to further improve the model. First, we would explore adding some preprocessing methods for face images, such as image sharpening, extended face alignment, and frontalization. Second, in terms of the model architecture, we may consider trying to combine layer functionalities such as with locally-connected layers or pooling layers.

ACKNOWLEDGMENT

This work was funded by the Michigan Aerospace Center for Simulations. We are grateful for our colleagues and the support of Central Michigan University.

REFERENCES

- [1] B. F. M. Cuneo, "22q11.2 deletion syndrome: Digeorge, velocardio-facial, and conotruncal anomaly face syndromes," *Current Opinion in Pediatrics*, vol. 13, 2001.
- [2] P. Kruszka, Y. A. Addissie, D. E. McGinn, A. R. Porras, E. Biggs, and M. Share. . . , "22q11.2 deletion syndrome in diverse populations," in *American Journal of Medical Genetics Part A*, 2017; 173 (4): 879 DOI: 10.1002/ajmg.a.38199.
- [3] Y. li Liu, W. Yan, and B. Hu, "Resistance to facial recognition payment in china: The influence of privacy-related factors," *Telecommunications Policy*, vol. 45, no. 5, p. 102155, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0308596121000598>
- [4] I. Olade, H.-n. Liang, and C. Fleming, "A review of multi-modal facial biometric authentication methods in mobile devices and their application in head mounted displays," in *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (Smart-World/SCALCOM/UIC/ATC/CBDCoM/IOP/SCI)*, 2018, pp. 1997–2004.
- [5] T. Zhu and L. Wang, "Feasibility study of a new security verification process based on face recognition technology at airport," *Journal of Physics: Conference Series*, vol. 1510, no. 1, p. 012025, mar 2020. [Online]. Available: <https://doi.org/10.1088/1742-6596/1510/1/012025>
- [6] S. Yamanaka and V. Moshnyaga, "New method for medical intake detection by kinect," in *2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2018, pp. 218–221.
- [7] M. Andrejevic and N. Selwyn, "Facial recognition technology in schools: critical questions and concerns," *Learning, Media and Technology*, vol. 45, no. 2, pp. 115–128, 2020. [Online]. Available: <https://doi.org/10.1080/17439884.2020.1686014>
- [8] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.
- [9] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *BMVC*, 2013.
- [10] T. Simonite, "Facebook creates software that matches faces almost as well as you do," *MIT Technology Review*, 2014.
- [11] C. NEWS, "Facebook's deepface shows serious facial recognition skills," *CBS NEWS*, 2014.
- [12] R. Bandom, "Why facebook is beating the fbi at facial recognition," *The Verge*, 2014.
- [13] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.
- [14] S. I. Serengil and A. Ozpinar, "Lightface: A hybrid deep face recognition framework," in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2020, pp. 1–5.
- [15] L. Wang and D.-C. He, "Texture classification using texture spectrum," *Pattern Recognition*, vol. 23, no. 8, pp. 905–910, 1990. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/0031320390901358>
- [16] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proceedings of 12th International Conference on Pattern Recognition*, vol. 1, 1994, pp. 582–585 vol.1.
- [17] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 32–39.
- [18] K. S. do Prado, "Face recognition: Understanding lbph algorithm," *towards datascience*, 2017.
- [19] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.
- [20] W. Ouyang, X. Wang, X. Zeng, S. Qiu, P. Luo, Y. Tian, H. Li, S. Yang, Z. Wang, C.-C. Loy, and X. Tang, "Deepid-net: Deformable deep convolutional neural networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [21] songguangfan, "A detailed explanation of deepid-net," June 2020.
- [22] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [23] Y. Bengio, "Learning deep architectures for ai," *Foundations and Trends® in Machine Learning*, 2009.
- [24] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *ANIPS*, 2012.
- [25] M. Valueva, N. Nagornov, P. Lyakhov, G. Valuev, and N. Chervyakov, "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation," *Mathematics and Computers in Simulation*, vol. 177, pp. 232–243, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0378475420301580>
- [26] W. Zhang, K. Doi, M. L. Giger, Y. W. R. M. Nishikawa, and R. A. Schmidt, "Computerized detection of clustered microcalcifications in digital mammograms using a shift-invariant artificial neural network," *Medical Physics*, vol. 21, pp. 517–524, 1994.
- [27] W. Zhang, K. Itoh, J. Tanida, and Y. Ichioka, "Parallel distributed processing model with local space-invariant interconnections and its optical architecture," *Appl. Opt.*, vol. 29, no. 32, pp. 4790–4797, Nov 1990. [Online]. Available: <http://ao.osa.org/abstract.cfm?URI=ao-29-32-4790>
- [28] S. Lawrence, C. L. Giles, Ah Chung Tsoi, and A. D. Back, "Face recognition: a convolutional neural-network approach," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 98–113, 1997.
- [29] B. Kayalibay, G. Jensen, and P. V. D. Smagt, "Cnn-based segmentation of medical imaging data," *ArXiv*, vol. abs/1701.03056, 2017.
- [30] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing [review article]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55–75, 2018.
- [31] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 International Conference on Engineering and Technology (ICET)*, 2017, pp. 1–6.
- [32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [33] A. Hue, "Dense or convolutional neural network part 1 — architecture, geometry, performance," *Medium*, 2020.
- [34] L. Weng, "An overview of deep learning for curious people," June 2017.
- [35] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2518–2525.
- [36] L. Wolf, T. Hassner, and I. Maoz, "Face recognition in unconstrained videos with matched background similarity," in *CVPR 2011*, 2011, pp. 529–534.

Design of Smart IoT Device for Monitoring Short-term Exposure to Air Pollution Peaks

Eric Nizeyimana¹, Jimmy Nsenga², Ryosuke Shibasaki³, Damien Hanyurwimfura⁴, JunSeok Hwang⁵

ACEIoT, University of Rwanda, Brussels, Belgium^{1,2,4}

Department of Socio-Cultural Environmental Studies³

Spatial Information (Sensing, Simulation and Services), University of Tokyo, Tokyo, Japan³

Department of Technology Management, Economics, and Policy, Seoul National University (SNU), Seoul, South Korea⁵

Abstract—Air pollution spikes have been causing harm to human beings and the environment. Most exposure to Air pollution spikes has demonstrated a significant impact on mental health, especially children at an early age. That lead to suicide or depression. Previous research concentrated on air pollution in general. Existing monitoring systems do not consider Short-term air pollution peaks. This paper presents the co-design of the hardware and software for IoT to monitor air pollution spikes for a short duration in real-time monitoring. The system comprises two technologies like edge computing to capture short-term exposure and a mathematical model for distribution in analyzing the captured data. This system ensures the presence of the spikes start and end for each pollutant. Monte Carlo simulation has been used in this research to predict the next spike of each pollutant. Artificial Intelligent is used to analyze immutable data for a short term prediction. After the analysis, legislators based on intelligent contracts created using blockchain to reduce pollution based on its source.

Keywords—Short-duration air pollution peak/spike; real-time monitoring; short-term prediction; immutable data; blockchain; AI

I. INTRODUCTION

Air pollution is the silent, prolific and invisible killer from previous years [1]. Most existing systems for monitoring air pollution are measuring the long-term peaks. Instead, the research shows that the short-term peaks are perilous [2] and can lead to different diseases such as eye and adnexa [3], brain volume, cognitive decrements, dementia development [4], heart, chronic obstructive pulmonary disease (COPD), lung cancer, migraine, acute lower respiratory infections and stroke [5]. Air pollution is exposed to more than nine to ten children and is stunting their brains, affecting their health [6]. That leads to the problem of mental health, especially in children (brain cell inflammation). Short-term spikes of air pollution are the source of increased hospital visits for childhood psychiatric. And the research shows that children from low-income families are more affected, leading to an increase of 44% of those who visited the hospital with suicidal thoughts due to the spikes in the air pollution [7] [8].

The spikes of air pollution have more severe effects on the brain of children. The research shows that air pollution spikes can cause mental health, depression, and anxiety. It can lead to the children having a lower intelligence quotient (IQ), poorer memory, delaying their development, leaving women infertile earlier, Spikes affect brain chemistry differently; for example, industries and traffics may carry toxins using tiny

passageways and then enter directly into the brain [7]. In 2020, spikes increased higher than five years [9]. But until now, minor effects of short-term exposure to air pollution are known.

Spikes of air pollution are damaging the future generation of humans, and emitters do not condemn the creation of that harm. Governments have tried to prevent air pollution in general, but the research shows that none of them has been viewing the spikes as a dangerous and long-lasting killer of children. The research suggested that legislators should protect children's' exposure to air pollution to advance the initiative for their public health [10].

This research paper is composed of the following section: The section of background covers all literature reviews related to the monitoring system of Air Pollution. The following section is about the co-design of hardware and software for the prototype monitoring of the spikes. The following section is about performance analysis and simulation, where the analysis made for some data and tools used to do simulation and the results shown in that section. The last section is the conclusion, which summarizes the paper and the proposal for future research.

II. BACKGROUND

This section highlights the background of the existing systems in monitoring Air Pollution. The section deeply explains the previous research and enumerates some challenges that are still in this area that can be solved using this research.

WHO has put the Global Air quality guidelines in different years to prevent air pollution in general, but there are no measures taken for spikes [11]. Many people are victims of air pollution, and emitters of pollutants are not charged for anything because none knows the air pollution they produce. Some measures have been taken [11], but they do not regularize the correspondence of emitters and victims of spikes. When these spikes continue to be repetitive, they cause more problems of health [12].

Many people are living in big and small cities. Developing countries used to have high populations exposed to air pollution. It is also where most sources of pollutants are found [13]. Once the spikes appear in the cities, it affects a large population [14]. Spikes can appear anytime, so this may come from different sources. If the spikes are not monitored, they can affect the living of human beings, as explained above. Spikes

Partnership for Skills in Applied Sciences, Engineering and Technology (PASET) under the Regional Scholarship and Innovation Fund (RSIF).

occur in a short time, and most existing monitoring systems for air pollution cannot recognize their appearance. Children are the most affected, primarily their mental, leading to their future loss [7]. Emitters of spikes may not even know how they affect human beings' health because there are no systems to monitor these spikes [15].

Authorities are oriented toward monitoring air pollution in general [16] [17]. Instead, spikes are affecting the future generation and the population as well. The source of spikes allows the identification of emitters, and then authorities may take advanced majors accordingly. Victims of these spikes are more in danger once they are repetitive. That may lead to many unexpected severe health problems.

The author of [18], proposed a system that can monitor the spikes from air pollution and predict the next spike using road management data.

Air pollution spikes have a short lifetime, requiring monitoring in a smaller time resolution [19]. These spikes need to be monitored at each appearance not to affect the living. Spikes need particular ways of monitoring them that differ from the existing methods. They appear in a short time, and then they disappear. If they do not monitor their appearance, they mix with other collected results of pollution and then may result in the average instead of the over-level for pollutants. Spikes monitoring can enable counting all spikes passed, predicting the following occurring spikes in the system. Once they occur from their sources, these unexpected pollutants may damage many things because they do not prepare before. They do not last for an extended period, leading to the big mess of not monitoring them. Spikes generated due to some occurred events planned before, but no analysis of the effect may cause. Also, spikes may occur due to unexpected events from the environment or any other source without any prior planning of the event. Spikes need a real-time and a low-time resolution to react to the effects that may occur due to its presence [20] [21], sometimes to the loss of life [22].

Most existing systems for monitoring air pollution are based on cloud-centric architecture [23] [24]. These systems measure air pollution with long term exposure. The average of peaks for air pollution in a specific time is considered the result of a monitored place. That is because of allowing sensors to capture information during a specific time and wake up to send the data on the cloud, known as duty cycle mode (taking a long sleep period to save the battery energy). That is for saving battery life during the wireless communication mode of sending data to the cloud. In the design of the sensor node of IoT applications, battery life is one of the critical parameters to consider. The reporting of collected data to the remote centric-cloud architecture of air pollution has a low frequency for at least 1hour to extend the battery lifetime.

The centric-cloud architecture uses wireless communication for transferring data from sensors to the remote. That leads to high energy consumption, and the sensors are sleeping within a certain period of collecting data and storing them locally. That makes sensors monitor long-term average peaks instead of capturing all peaks [25]. That leads to the miss of monitoring short duration peaks. These spikes may

appear periodically or not. The air pollution peak average threshold may exceed for specific pollutants, and the system may not be aware of that unexpected change. No system can capture spikes for a short duration from the existing cloud-based systems.

Cloud-centric has failed to monitor spikes because of transferring data by waking up the sensor. The cloud-centric architecture collects data of air pollution using sensors at data gathering. It uses wireless communication to send the data to the cloud [26, 27]. Then, the system does the data management for the given application in the cloud. At the first phase of data gathering, sensors collect information related to air pollution. This information is transferred to the cloud through the wireless communication channel. This communication channel consumes high energy [28]. The last phase is data management, which analyses, processes, and stores data in the cloud. These data can be used to predict air quality [29].

The cloud-centric architecture also has a latency problem due to transferring the data after a specific time. These systems also take time to react to the processed data [30]. Predicting the possible air pollution event may take longer as data processing is based on the cloud, not on edge. There is a need for data transparency and trust, and this may be difficult for the data passing in the network without additional security measures.

Therefore, there is a need for an edge-centric system to monitor short-term peaks for air pollution. This paper helps to understand the design of the edge-centric smart sensor to monitor air pollution by waking up the sensor once the air pollution variation attained the given threshold.

There are not many several systems developed to monitor and predict air pollution spikes. Artificial Intelligence technology is the predictor developed by a new wireless company, and that system can predict the following levels of air pollution within an hour. This system uses AI for analyzing weather measurements, images of CCTV cameras, air pollution sensing devices, Bluetooth, and history readings. The system links the existing real-time data to predict the next coming hour for traffic jams and air pollution. This predictor is accurate at 97%. It has been tested for implementation in some cities like Wolverhampton [18]. It is excellent and friendly to the existing technologies, but it doesn't take pollutant data on the roads or nearby since it is linked to the load management system. And this leads to the lack of identifying the source of each pollutant and the quantity. The predictor predicts air pollution in general but doesn't identify spikes that come and may arrive.

Following the previous works that researchers have done, the existing systems only measure a few pollutants. Most monitor air pollution at the cloud-centric, leading to latency, security, cost, and control problems. The existing system woke up sensors periodically, leading to the danger of an unexpected increase of pollutants. Existing systems haven't mentioned the identification of spikes and the time stay based on their appearance level.

This research takes all six primary pollutants as explained by WHO in [11], and it can be installed and not based on historical readings.

III. HARDWARE AND SOFTWARE CO-DESIGN

A. Improve IoT Energy Management

In designing the embedded system for performing a real-time environment, there is the issue of the increase of power dissipation. IoT hardware design increases power dissipation from the real-time application and the device for the best performance. The problem was created during the deployment of the number of transistors comparable with the power consumption. There are two causes of power dissipation in designing lower-power IoT systems. The first one is when the power dissipation for each transistor increases with impact to the increase of density gate, which implies the increase of power density for the whole system. The second is the increase of the frequency of IoT systems for better performance.

The power dissipation problem is improving IoT energy management by waking up the device using analogue interrupts.

Energy management is still a crucial problem in today's sensors [31]. There is a need to continuously allow the sensor to stay in energy-saving deep sleep mode to solve that issue. The system needs to wake up on measurement appearance with low energy consumption. Since the energy consumption implies a decrease in battery life, the system should monitor sensors connected to use little energy.

The system stays asleep most of the time and only wakes up for the threshold's quick and effective measurement. The CPU of the system uses much energy by comparison to the rest of the other parts. That means that reducing the CPU system's busy time is the best way to reduce consumption energy.

Fig. 1, the system wakes up periodically to detect events, making the CPU continuously active. The sensors capture data from the environment and send these data for processing. The analogue event creates a signal which is transformed directly to a digital signal.

Fig. 2 gives the intention to look at the signal after using analogue interrupts. The sensors can record all analogue events, and once the threshold passes, the system wakes up for recording and processing.

There are two options for using analogue interrupts: The first is ADC wake up and the second is an external op-amp-based voltage comparator.

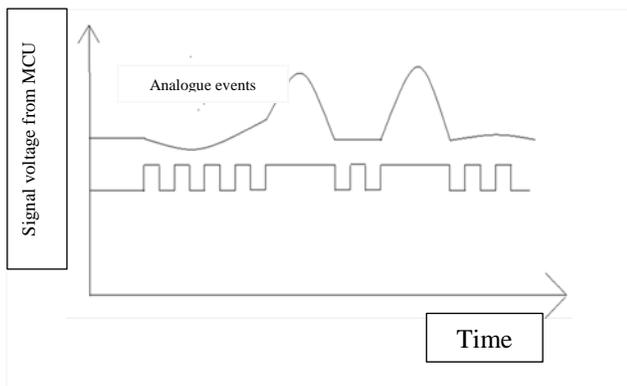


Fig. 1. System Wakes up Periodically to Detect Events.

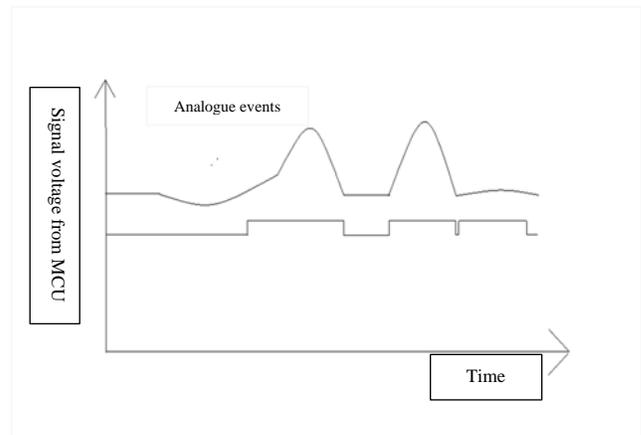


Fig. 2. System Wakes up on the Threshold.

1) *ADC Wake up:* Sensors convert analogue measurements into electronic signals. The ADC (Analog to Digital Converter) converts the produced analogue signal to a digital signal using the frequency sampling mode based on the Nyquist theorem. Interrupts that are alerting electronic signals are sent to the Microcontroller Unit (MCU) processor, which may come as an external part of the internal peripheral or external one.

The below is an ADC component designed using the Proteus simulator. The ADC is composed of one ADC0804 Integrated Circuit, eight LEDs, one resistor of 1k, one variable resistor or potentiometer, one push button, one wero board, one nonpolar capacitor with 150pf and some jumper wires. This proteus simulation designs system that switches ON and OFF based on the voltage once the input exceeds the threshold. Then, it helps to monitor all events that come and exceed the threshold. That is useful in air pollution monitoring based on spikes only.

Fig. 3 uses the button to switch off the ADC. The button is activated by the measurement to the node sensor for air pollution. The input signal is an analogue signal generating the output that can switch on LEDs.

Fig. 4 shows that the data are generated from the physical environment, and then the sensor accepts these analogue measurements in the form of analogue signals. The analogue signal transforms into a digital signal and is then sent to the processor. At the input, the environment creates a signal using physical quantity. Then sensor takes that signal and makes it in the presence of an electrical signal. The signal is in analogue form and needs to transform into digital form, and then using the ADC tool; it generates the digital signal used to monitor air pollution using the given threshold.

This paper applies the sampling of analogue signals, and the system acts based on the threshold. That should be done using ADC to trigger timers precisely. That uses many MCU resources, which leads to high-power consumption since timers must be active to perform ADC. Another methodology is not to use timers and allow signals to be monitored continuously by the ADC, which consumes a high-power consumption.

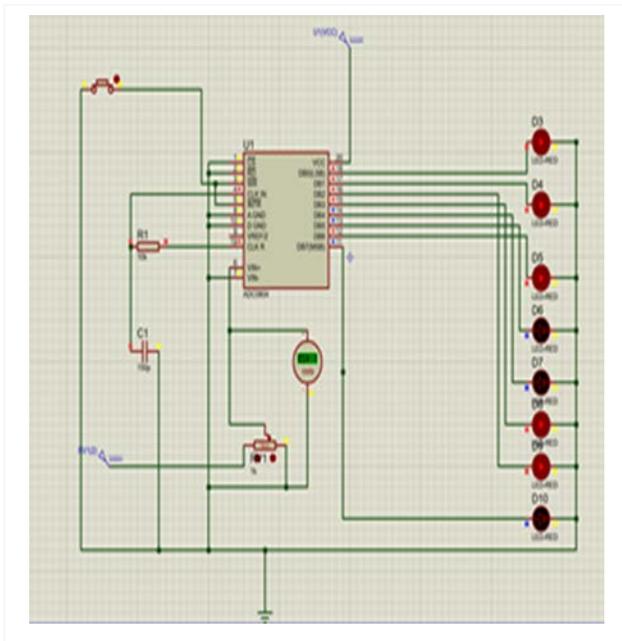


Fig. 3. ADC Circuit from Proteus Simulation.

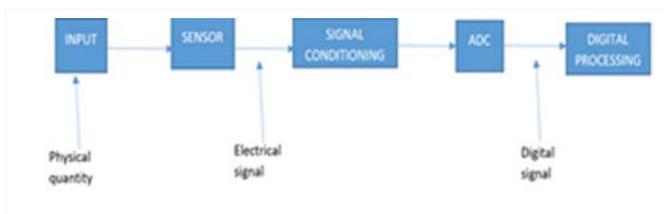


Fig. 4. Measurement to Signal Processing.

The solution to be adapted that may not consume huge amounts of power is to integrate ADC in the MCU without dependency on the CPU (Central Processing Unit). That allows the CPU to disable all clocks except the one of ADC. Then the ADC wakes up the CPU and other parts of the MCU by using the logical conditions. The ADC creates interrupts to wake up the rest of the system by referencing the configured threshold.

The ADC wake up uses the voltage comparator to activate the CPU and the system. The reference voltage V_{ref} is compared to the input voltage V_{in} for deciding to wake up the system or not. The V_{out} is in digital mode, and from there, the decision to wake up the system is taken. Since that is on the sensor by detecting the measurement of the event to wake up the system, there is the optimality of this strategy because no loss of data appears and the optimization of the sleeping time.

2) *External op-amp based voltage comparator*: The other way to wake up the system is to use an external operational amplifier based on the voltage comparator. This way requires extra resources to add to the system. And adding this wake-up circuit to the sensor node decreases the average power consumption, but also it may create the loss of information since the original signal was amplified.

The external op-amp compares one analogue voltage level to another and generates output based on the comparison. It

detects the voltage from measurements and then switches from the sleeping mode of the system to an active mode. The switching time of the op-amp voltage comparator slows the system even though it operates on analogue voltage.

The op-amp voltage comparator uses input, amplification, and output terminals. It uses negative feedback voltage, which leads to compensation capacitance to prevent oscillation in that integrated circuit. That creates an inside power dissipation, which may increase the temperature for the chip and the self-heating.

The operational amplifier voltage comparator may present an error voltage called the input offset voltage caused by the characteristics of transistors of each terminal or by the input bias current.

Fig. 5 shows the graph of the op-amp voltage comparator with its five terminals.

Therefore, based on the above comparison of ADC wake up and external op-amp voltage comparator wake up, this paper suggests using ADC wake up to activate the system from the sleeping mode to active mode. The measurements are taken from the environment and create analogue input to the sensor node, and that analogue input changes to an analogue signal with a certain amount of voltage. Then the analogue signal needs to transform into a digital signal using an ADC converter, and during that conversation, the ADC decides if it wakes up the whole system based on the comparison of the input voltage and the reference voltage. The system needs to identify the starting voltage above the threshold voltage and record these values. The following subsection explained how the system used to pick these signals that attained the threshold.

B. Peak Digital Signal Processing

Measurements captured from the environment need to be processed and analyzed. Once the ADC wakes up, it gets an analogue signal and converts it to a digital signal, quickly processing scientifically. The system is woken up based on the data that exceed the predefined threshold, allowing the system to record that event.

This research is working on air pollution spikes. These spikes are identified based on the minimum predetermined value of pollutants. WHO has defined each pollutant's threshold as shown in Table 1. These values are measured in micrograms per cubic meter.

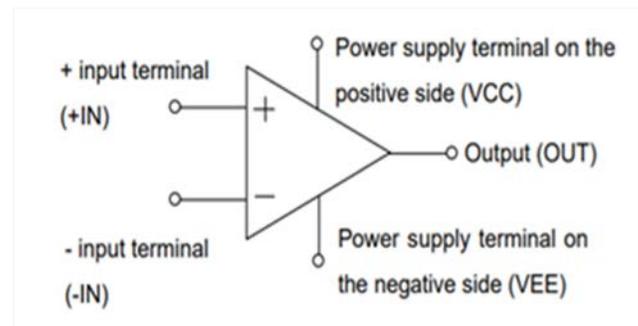


Fig. 5. Op-amp Voltage Comparator.

TABLE I. AIR POLLUTANTS

Pollutant name	Minimum Concentration ($\mu\text{g}/\text{m}^3$)
Particulate Matter $\text{PM}_{2.5}$	35
Particulate Matter PM_{10}	70
Carbone monoxide/ Carbone dioxide (CO/ CO_2)	1000
Nitrate Oxide (NO_x)	80
Sulfur of Oxide (SO_x)	50
Ozone (O_3)	120

This research detects a peak in a signal and measures its position, height, width, and/or area. When the sensor node identifies the signal that exceeds a threshold value of any type of pollutant, the system starts to record the event, and when it attains the peak, it starts to decrease, going to the value which should always be less than the threshold. The first derivative of the peak is applied to downward-going zero-crossing (threshold) at the maximum of the peak. Since the signal may have noise from measurement due to the environment, this can lead to false zero-crossing. Therefore, the smooth technique can detect only the desired peaks and ignore peaks that are too small, too wide, or too narrow.

Once they become high frequency, air pollution spikes (peaks) cause mental health problems that can lead to hypertension, suicide, and heart diseases. These peaks imply the concentration of pollutants in respect of the given time. If not reduced, that concentration cause health problems compared to normal pollutants that don't pass the threshold.

The input signal is taken in the window size measured based on the length of the signal above the threshold. The height of that signal is also identified.

The digital signal processing from ADC is set low or high. We examine all signals with high since they are above the threshold. Signals which are less than the threshold are identified as low. Then finding the peaks in the given signal that we can call X describes all points above the threshold. Each peak has its amplitude or the height of the signal.

Fig. 6 is for peak detection of digital signals with a height of 1 and the length of the period of 500 microseconds.

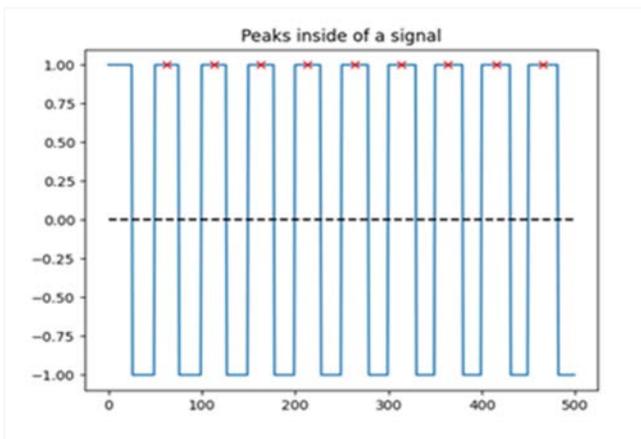


Fig. 6. Peaks Detection.

From the input signal of the sensor node, ADC wakes up activates the system once the peak has appeared. The probability of obtaining a peak at a particular point t_i of the input signal from the environment depends on its incoming voltage and the given standard threshold voltage t_v .

$$P(\text{Spike at } t_i | \text{Spike at } t_v) = \int_{v \in C_i} G[V(t)]$$

With C_i , the set of possible voltages can rise in the interval i . V presents the voltage signal. And $G[V(t)]$ is the Gaussian distribution.

Once peaks are detected from the given interval time and their location, the system is applied to collect these peaks. These peaks are stored in the array of integers for future analysis using Machine Learning. The system can analyze the collected peaks for notifying the appearance of pollutants and the expectation of the next peak.

IV. PERFORMANCE ANALYSIS / SIMULATION

A. Distribution Patterns of Air Pollution Spikes

Air pollution spikes are coming from the increase of unexpected pollution generated by emitters. All these spikes are from the environment and can be distributed in the atmosphere. The system for monitoring spikes can capture distributed pollutants either directly or indirectly from the source.

Direct spikes are captured by the system and then analyzed without adding the environment to it and for example, having the sensor node connected to the place generates pollutants. The indirect spikes are those that pass in the environment and meet with other pollutants before being measured by the system.

The system accepts measurement in any of three patterns or their combination. Spikes can be presented to the system either: uniform random and/or clumped.

Uniform spikes are these spikes that come within a given period. They occur periodically in the system. These spikes are easy to predict the next peak. These peaks can appear in different sizes and densities.

Random spikes are these peaks that are entered into the system randomly. These peaks can be very harmful since they are not easily predictable. This research suggests using a certain period to analyze all peaks appearing, and then using Machine learning, and it can predict the next peak.

Clumped spikes are predicted or unpredicted peaks with a heavy density. These peaks are perilous, and they need profound observations to analyze their prediction.

The nature of the air pollution environment can have all these above patterns of distribution of spikes. All these spikes are based on the period to predict the next appearance of the peak.

The distribution model has been used to predict the next within a specific period. Since the generated signal from the ADC wakes up converter is a digital signal, it has discrete values. The Poisson distribution is used in this paper to model the arrival rate of spikes in a specific fixed interval of time.

The performance parameters are based on the mean of signals in a period λ and the number of spikes k .

Let λ be the parameter greater than 0 and let distribution $k = 1, 2, 3, \dots, n$ be the appearance of spikes in the input signal to the sensor node; in other words, k is presenting a discrete random variable counted. The probability density function (*pdf*) is used to specify the random variable's probability being in the range of the values.

Then the *pdf* that a Poisson random variable X with the mean λ is equal to a given by the formula.

$$pdf = P(X = a) = \frac{\lambda k e^{-\lambda}}{k!}$$

Where e is a constant approximately to 2.71828.

The *pdf* gives the probability of getting spikes each time by using the mean of the spikes and the number of spikes counted.

The system can identify spikes and predict finding peaks in the given time. Most existing systems for monitoring air pollution are using cloud-centric duty cycle mode. The sensor collects data related to air pollution while it is in sleeping mode to save the battery's lifetime or the harvested power for energy conservation. The sensor only wakes up after a specific period to transmit collected data to the cloud-centric for the analysis.

The duty cycle D can be defined as a ratio of pulse width (PW), a busy time and the total period of T of the signal and then expressed in percentage.

$$D = \frac{PW}{T} \times 100$$

Therefore the 60% duty cycle means that the signal is on 60% of the time but off 40%. That implies the power consumption in recording data. There is a need for much energy during the transmission of the data to the cloud-centric server. This energy consumption reduces the sensor battery lifetime or the harvest power storage of energy.

The solution of optimizing the use of sensor energy or harvest power is edge centric HW/SW Codesign smart sensor. That allows the analogue interrupt to wake up the sensor once the coming signal voltage is higher than the threshold voltage. Only the system is active in collecting spikes for quick analysis of data. Once the spikes are over, the intelligent sensor goes back to sleep mode.

The above Fig. 7 describes the edge centric HW/SW co-design system. The measurement from the environment is those pollutants CO, CO₂, PM_{2.5}, PM₁₀, SO_x and NO_x. Once one of these pollutants sends the value more significant than the threshold defined in the table [1], the sensor node will send all signals to ADC, which wakes up the processor, memory, protocol for analyzing the coming signal since it is a spike. The power unit is there to empower each device. That reduces the power consumption since it is waking up for recording spikes.

In edge centric HW/SW co-design, data analysis is performed locally, and there is no transmission energy of short periodical time. Only transmission can be once in a while for further analysis. The edge-centric smart sensor is a real-time monitoring system for air pollution spikes and can react to its

appearance. From those spikes, it does analysis locally, and the decision is taken quickly.

The cloud-centric is still needed to analyze big data collected by sensors, while edge centric can be considered an operator of instant data. At the edge, centric analytical tools and AI tools are nearest the system, implying operational efficiency. The security and privacy are strong at the edge centric smart sensor. This system is reliable. Since one node can go down and is unreachable, the other system parts continue to operate. The speed of data at edge computing implies analytical, computational resources to the end-users, bringing quick responses and applications.

B. Performance Metrics

The energy consumption at the edge centric HW/SW co-design smart sensor and cloud-centric can be distinguished in the below metrics:

- Throughput: output at the edge is generated in real-time while data is transmitted in the cloud, which consumes much energy.
- Collecting data: at this stage, edge computing uses to wake up only during the collection of spikes while the cloud uses a duty cycle which implies power consumption.
- Processing: at computing only, the system wakes up on the threshold, while the CPU and other parts of the system operate periodically for cloud-centric—the probability of identifying the subsequent spikes at edge computing within the length of the interval of spikes.

From Table 2, the designed system performs better in all performing metrics. The designed system can save energy consumption at 40% on the throughput metric since the existing systems (mostly cloud computing systems) are using a duty cycle. On the second metric of data collection, the designed system woke up on the appearance of spikes, and that can lead to quickly identifying the spike while the existing systems wake up periodically and can miss some spikes, which may be dangerous. Lastly, the processing metric is so quick at the designed system while existing systems take enough time to process and predict the next appearance of the spike for air pollution.

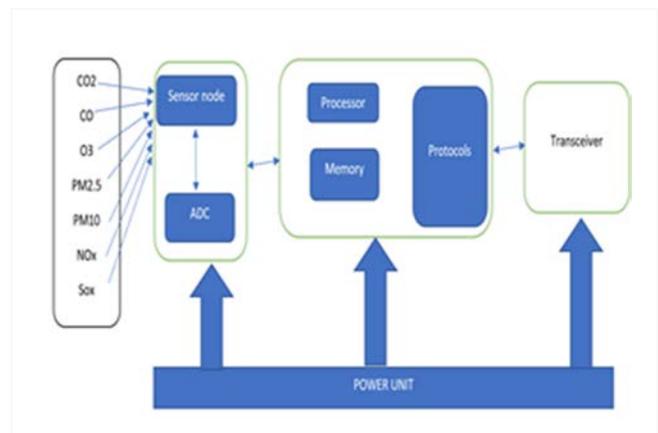


Fig. 7. Edge-Centric HW/SW Codesign System.

TABLE II. COMPARISON OF THE EXISTING SYSTEM AND OUR DESIRED SYSTEM

	Throughput	Data collection	Processing / Performance
Designed system	Save energy consumption at 40%	Spikes to wake up the system	Quick processing and prediction
Existing systems [32, 33]	Energy consumption	Periodically wake up	It takes periodically time to predict the next spike.

Peaks can be harmful to human beings, and there is a need to monitor them. For cloud computing, some peaks may be lost during the sleep mode of the sensor at the sleeping mode of the sensor. At the sleeping mode of the sensor, all peaks arrive and can be combined with the whole signal for presenting the mean of the whole period. On edge centric, the system is woken up by spikes in the input signal.

C. Monte - Carlo Simulation

This research uses Monte-Carlo simulation as a mathematical technique used to estimate the probability of possible outcomes in a process that cannot be predicted due to its uncertain appearance.

It is based on making a computational algorithm to find the numerical results of repeated random sampling. These uncertainty events can be predicted and forecasted using the Monte-Carlo technique to model them.

This research uses Monte-Carlo simulation to predict the subsequent spikes using their probabilities of occurring. As our data are stored discretely after the ADC converter, we estimate the probability of occurring in a specific period.

Let's use the same example of PM2.5 for its Poisson distribution, which was 7.1%. Then that means in the interval of a period there is a 7.1% probability to find the spikes of PM2.5. The figure below is for particulate matter data; and it uses the synthetic data generated mostly.generate. The mean λ of the data is 29, and the appearance of spikes is 27. Then the probability of getting the spikes is 0.071.

The performance of the proposed system has good accuracy since it can identify each appearance of the spike of each pollutant. Most other existing systems measure air pollution in general and are not specific on spikes prediction of each pollutant.

The signal length is sampled at 30 samples for the whole period, and the probability of finding the next peak of PM2.5 is 7.1%, as shown in Fig. 8.

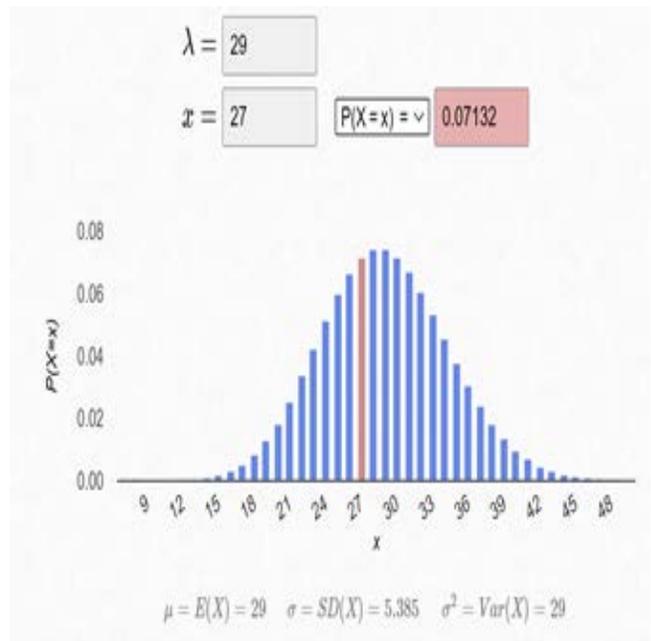


Fig. 8. Probability of following Peaks for PM2.5.

V. CONCLUSION

This research designs a hardware-software smart IoT device to monitor short-term exposure to air pollution peaks. Spikes are causing significant health problems, especially children, leading to mental problems, suicide, stroke, heart diseases, lung, etc. This research improves the IoT energy by waking up the system through the appearance of digital signals using ADC wake up. The designed system performs better than the existing system through the performance metrics, as explained in Table 2. The paper explained the finding of peaks that are stored in the array. The mathematical model was generated using Poisson distribution to find the appearance of peaks. Monte-Carlo has been introduced to predict the next coming peak. The prediction showed that PM2.5 could be predicted at a 7.1% probability of the spikes to appear. This probability is high since it showed that in the appearance of spikes, there should be a 7.1% of PM2.5.

The system and authorities analyze the collected peaks to compensate for the peak emitters. As peaks are dangerous to health, there should be a proposal of fining people accordingly if they exceed the threshold. The hardware-software co-design generates a dataset of spikes signature that will be used by machine learning for future research. In future works, this research will be oriented on the security of the data transmitted across the network using blockchain.

REFERENCES

- [1] U. Nations, "Air pollution, the 'silent killer' that claims seven million lives a year: rights council hears.," <https://news.un.org/en/story/2019/03/1034031>, 2019.
- [2] Eric Nizeyimana, Damien Hanyurwimfura, Ryosuke Shibasaki, Jimmy Nsenga, "Design of a decentralized and predictive real-time framework for air pollution spikes monitoring," in 2021 IEEE 6th International Conference on Cloud Computing and Big Data Analytics (ICCCBDA), Chengdu, China, 2021.
- [3] Jie Song, Yue Liu, Mengxue Lu, Zhen An, Jianguo Lu, Ling Chao, Liheng Zheng, Juan Li, Sanqiao Yao, Weidong Wu, Dongqun Xu, "Short-term exposure to nitrogen dioxide pollution and the risk of eye and adnexa diseases in Xinxiang, China," *Atmospheric Environm*, vol. 218, p. 11700, 2019.
- [4] Gao, Xu and Coull, Brent and Lin, Xihong and Vokonas, Pantel and Spiro, Avron and Hou, Lifang and Schwartz, Joel and Baccarelli, Andrea A, " Short-term air pollution, cognitive performance and nonsteroidal anti-inflammatory drug use in the Veterans Affairs Normative Aging Study," *Nature Aging*, vol. 1, pp. 430-437, 2021.
- [5] Ali, Muhammad Ubaid and Yu, Yangmei and Yousaf, Balal and Munir, Mehr Ahmed Mujtaba and Ullah, Sami and Zheng, Chunmiao and Kuang, Xingxing and Wong, Ming Hung, "Health impacts of indoor air pollution from household solid fuel on children and women," *Journal of Hazardous Materials*, vol. 416, p. 126127, 2021.
- [6] U. Nations, "More than nine in ten children exposed to deadly air pollution," 2018.
- [7] S. Nickerson, "Even Small Spikes in Air Pollution Can Threaten Children's Mental Health, Research Suggests," 2019.
- [8] V. Sreenivasan, *Air Pollution and Child Mental Health*, 2021.
- [9] NATALIE RAHHAL, SAM BLANCHARD,, "Spikes in air pollution 'trigger rises in the number of children needing emergency hospital treatment for anxiety or suicidal thoughts," 2019.
- [10] Lisa Potter, "Air pollution spikes reduce test scores," 2020.
- [11] L. Potter, "New WHO Global Air Quality Guidelines aim to save millions of lives from air pollution," 2021.
- [12] UIAmin, Riaz and Akram, Muhammad and Ullah, Najeeb and Ashraf, Muhammad and Malik, Abdul Sattar, "IoT Enabled Air Quality Monitoring for Health-Aware Commuting Recommendation in Smart Cities," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, pp. 279 - 285, 2020.
- [13] Shaddick, G and Thomas, ML and Mudu, P and Ruggeri, G and Gumy, S., "Half the world's population are exposed to increasing air pollution," *NPJ Climate and Atmospheric Science*, vol. 3, pp. 1-5, 2020.
- [14] Mullins, Jamie and Bharadwaj, Prashant, "Effects of Short-Term Measures to Curb Air Pollution: Evidence from Santiago, Chile," *American Journal of Agricultural Economics*, vol. 97, no. 4, pp. 1107-1134, 2015.
- [15] S. Mayor, "Spikes in air pollution raise heart risk as much as sustained exposure, study suggests," *British Medical Journal Publishing Group*, 2018.
- [16] Cocozza, Claudia and Alterio, Edoardo and Bachmann, Olivier and Guillong, Marcel and Sitzia, Tommaso and Cherubini, Paolo, "Monitoring air pollution close to a cement plant and in a multi-source industrial area through tree-ring analysis," *Environmental Science and Pollution Research*, pp. 1-11, 2021.
- [17] Aziz, Zena A Aziz and Ameen, Siddeeq Y Ameen, "Air pollution monitoring using wireless sensor networks," *Journal of Information Technology and Informatics*, vol. 1, no. 1, pp. 20-25, 2021.
- [18] P. Neill, "Artificial Intelligence technology to predict air pollution spikes," 2020.
- [19] Subramanian, R and Kagabo, Abdou Safari and Baharane, Val{e}rien and Guhirwa, Sandrine and Sindayigaya, Claver and Malings, Carl and Williams, Nathan J and Kalisa, Egide and Li, Haofan and Adams, Peter and others, "Air pollution in Kigali, Rwanda: spatial and temporal variability, source contributions, and the impact of car-free Sundays," *Clean Air Journal*, vol. 30, no. 2, pp. 1-15, 2020.
- [20] Johnson, Stacy A and Mendoza, Daniel and Zhang, Yue and Pirozzi, Cheryl S, "Effects of Short-Term Air Pollution Exposure on Venous Thromboembolism: A Case-Crossover Study," *Annals of the American Thoracic Society*, 2021.
- [21] Lammers, Ariana and Neerinx, Anne H and Vijverberg, Susanne JH and Longo, Cristina and Janssen, Nicole AH and Boere, A John F and Brinkman, Paul and Cassee, Flemming R and van der Zee, Anke H Maitland, "The Impact of Short-Term Exposure to Air Pollution on the Exhaled Breath of Healthy Adults," *Sensors*, vol. 21, no. 7, p. 2518, 2021.
- [22] Scortichini, Matteo and De Sario, Manuela and De'Donato, Francesca K and Davoli, Marina and Michelozzi, Paola and Stafoggia, Massimo, "Short-term effects of heat on mortality and effect modification by air pollution in 25 Italian cities," *International journal of environmental research and public health*, vol. 15, no. 8, p. 1771, 2018.
- [23] Senthilkumar, R and Venkatakrishnan, P and Balaji, N, "Intelligent based novel embedded system based IoT enabled air pollution monitoring system," *Microprocessors and Microsystems*, vol. 77, p. 103172, 2020.
- [24] Toma, Cristian and Alexandru, Andrei and Popa, Marius and Zamfiroiu, Alin, "IoT Solution for Smart Cities' Pollution Monitoring and the Security Challenges," *Sensors*, vol. 19, no. 15, p. 3401, 2019.
- [25] Liu, Jun and Yin, Hao and Tang, Xiao and Zhu, Tong and Zhang, Qiang and Liu, Zhu and Tang, XiaoLong and Yi, HongHong, "Transition in air pollution, disease burden and health cost in China: A comparative study of long-term and short-term exposure," *Environmental Pollution*, vol. 277, p. 116770, 2021.
- [26] Kolumban-Antal, Gyorgy and Lasak, Vladko and Bogdan, Razvan and Groza, Bogdan, "A secure and portable multi-sensor module for distributed air pollution monitoring," *Sensor*, vol. 20, no. 2, p. 403, 2020.
- [27] Zakaria, Nurul Azma and Abidin, Zaheera Zainal and Harum, Norharyati and Hau, Low Chen and Ali, Nabeel Salih and Jafar, Fairul Azni, "Wireless Internet of Things-based Air Quality Device for Smart Pollution Monitoring," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, pp. 65--69, 2018.
- [28] WLin, Hang and Chen, Huangxin and Zhang, Lin and Luo, Youjia and Shi, Yi and Zou, Wenjie, "Energy consumption, air pollution, and public health in China: based on the Two-Stage Dynamic Undesirable DEA model," *Air Quality, Atmosphere & health*, pp. 1-16, 2021.
- [29] Delgado, Alexi and Acuna, Ramiro Ricardo Maque and Carbajal, Chiara, "Air Quality Prediction (PM2. 5 and PM10) at the Upper Hunter Town-Muswellbrook using the Long-Short-Term Memory Method," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 318-322, 2020.
- [30] Wang, Yan and Liu, Xiaotian and Chen, Gongbo and Tu, Runqi and Abdulai, Tanko and Qiao, Dou and Dong, Xiaokang and Luo, Zhicheng and Wang, Yikang and Li, Ruiying and others, "Association of long-term exposure to ambient air pollutants with prolonged sleep latency: The Henan Rural Cohort Study," *Environmental Research*, vol. 191, p. 110116, 2020.
- [31] Sadeeq, Mohammed AM and Zeebaree, Subhi, "Energy management for internet of things via distributed systems," *Journal of Applied Science and Technology Trends*, vol. 2, no. 2, pp. 59-71, 2021.
- [32] Okokpujie, Kennedy O and Noma-Osaghae, Etinosa and Odusami, Modupe and John, SN and Oluga, Oluwatosin, "A smart air pollution monitoring system," *International Journal of Civil Engineering and Technology (IJCIET)*, vol. 9, no. 9, pp. 799 - 809, 2018.
- [33] Shitharth, S and Manimala, Ms Korukonda and Bhavani, Ms VV and Nalluri, Mr Srikanth, "Real Time Analysis of Air Pollution Level in Metropolitan Cities by Adopting Cloud Computing based Pollution Control Monitoring System using Nano Sensors," *Solid State Technology*, pp. 1031-1045, 2020.

Robust Facial Recognition System using One Shot Multispectral Filter Array Acquisition System

M. Eléonore Elvire HOUSSOU¹, A. Tidjani SANDA MAHAMA², Pierre GOUTON³, Guy DEGLA⁴
ImVIA, University of Bourgogne Franche-Comté, Dijon France^{1,2,3}
IMSP, University Abomey-Calavi, Dangbo Benin^{1,2,4}

Abstract—Face recognition in the visible and Near Infrared range has received a lot of attention in recent years. The current Multispectral (MS) imaging systems used for facial recognition are based on multiple cameras having multiple sensors. These acquisition systems are normally slow because they take one MS image in several shots, which makes them unable to acquire images in real time and to capture moving scenes. On the other hand, currently there are snapshot multispectral imaging systems which integrate a single sensor with Multispectral Filter Arrays (MSFA) allow having at each acquisition an image on several spectra. These systems drastically reduce image acquisition time and are able to capture moving scenes in real time. This paper proposes a study of robust facial recognition using Multispectral Filter Array acquisition system. For this goal, a MSFA one-shot camera was used to collect the images and a robust facial recognition method based on Fast Discrete Curvelet Transform and Convolutional Neural Network is proposed. This camera covers the spectral range from 650 nm to 950 nm. A comparison of the facial recognition system using Multispectral Filter Arrays camera is made with those that using multiple cameras. Experimental results proved that face recognition systems whose acquisition systems are designed using MSFA perform more efficiently with an accuracy of 100%.

Keywords—Multispectral image database; multispectral imaging; multispectral filter array (MSFA); one-shot camera; facial recognition system

I. INTRODUCTION

The Biometric system is defined as an automatic measurement system based on the recognition of physiological and / or behavioral characteristics specific to a person. It is characterized by its uniqueness, its public nature, and its performance. There are several biometric modalities namely fingerprint, palm sign, face, iris, retina, DNA, voice etc.

Facial recognition is one of the widely used biometric identification methods because face is easy to capture in a controlled or not controlled environment, and in a cooperative or non-cooperative manner [1] [2]. Facial recognition systems performance depends on the electromagnetic spectra in which face images have been acquired. Facial recognition based on the visible spectrum generally use texture characteristics. Its performance is affected by light, occlusions, and pose variations. Infrared spectrum has several advantages over the visible spectrum; it is not perceptible to the human eye and at the same time, less sensitive to variations in light[3] [4]. The infrared spectrum is subdivided into near infrared spectrum (770-1400 nm), Short Wavelength Infrared (1,4–3 μm), mid-

wave infrared spectrum (3 – 8 μm) and thermal infrared spectrum (8 - 15 μm). There have been reported some research showing that in environments with uncontrolled illumination the NIR approach remarkably has higher performance in comparison to VIS approach in the extraction of information in different aspects such as appearance and structure [5]. The use of visible and near infrared spectra in facial recognition combines the benefits of both spectra and improves the performance of facial identification systems. Face recognition in the visible and infrared range has received a lot of attention in recent years. A multispectral recognition system is a system using images acquired on 3 to 10 spectral bands. Each of the images acquired in a given band contains specific information that is very important and useful in facial recognition.

MS imaging [6] systems can be broadly grouped into three categories: multi-cameras systems, single-camera and multi-shot systems, and single-camera one-shot. The multispectral systems based on the first category consist of several cameras, at least one per inference filter (or spectral band filter), multispectral systems using a single camera and several shots are made up of several image sensors each equipped with a narrow bandwidth wavelength filter, which makes them heavy, bulky, energy-intensive and very expensive. The last category that of single camera one shot, uses a single sensor with Spectral Filter Array (SFA) or Multispectral Filter Array (MSFA) to acquire a single image on multiple spectral band simultaneously. These imaging systems are very fast and operate in real time.

Most current MS imaging facial recognition systems use acquisition systems consisting of multiple cameras or a single camera with multiple sensors. Considering the real time operations and benefits of these one-shot multispectral cameras, the research has been oriented towards facial recognition using multispectral images acquired with Multispectral Filter Array camera. This paper proposes a robust facial recognition system using MS image dataset collected with MSFA one shot camera that covers the visible and near infrared spectrum. Fast Discrete Curvelet Transform(FDCT)[7] ,VGG19 [8] and ResNet 101[9] convolutional neural networks have been used to develop recognition end.

This paper is organized as follows: Section 2 focuses on different MS facial recognition systems used in the literature. Section 3 describes our Multispectral Filter Array one shot acquisition system and the method. The experimental results are reported in Section 4. Section 5 is dedicated to the discussion and the conclusion is presented in the last section.

II. RELATED WORK

The use of multiple spectral bands improves the performance of facial recognition system, this explains the interest of researchers on MS facial recognition in recent years [5] [10] [11]. A face recognition system can be represented by four main modules: capture, feature extraction, matching and decision. The capture module is mainly based on image acquisition system that enables the acquisition of images. The acquisition system consists of one or several cameras. The feature extraction module takes the acquired images and extracts only the relevant information in order to build a new data representation. The matching module compares the set of extracted features with the features of those images stored by the system in the database during enrollment. The decision module verifies the identity asserted by a user based on the degree of similarity between the extracted features and the ones from the database. The following figure (Fig. 1) illustrates the architecture of a face recognition system.

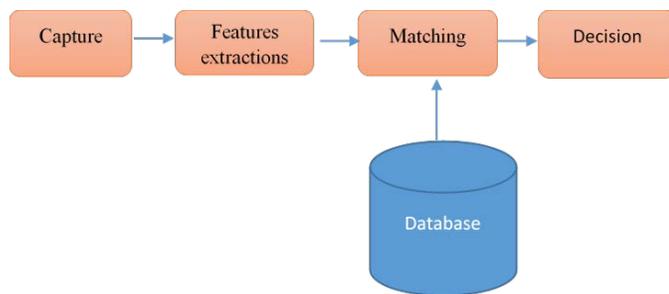


Fig. 1. Architecture of a Facial Recognition System.

There are several facial recognition systems operating in the visible and Near InfraRed range, but in the most of the cases images were not acquired in real time.

Aboud et al. [12] proposed a face recognition system using Fusion of Multispectral Imaging to overcome the limitations of visible facial recognition. This recognition system merges features from visible, near infrared and thermal infrared images. This system did not acquire images in the capture module but used images from the Carl database for the other modules. Carl Face images Database contains visible, NIR and thermal images of 41 persons. This database had been collected with two cameras: customized Logitech Quickcam messenger E2500 with a Silicon based CMOS image sensor for near infrared image and thermal camera TESTO 880-3 (incorporating an uncooled detector with a spectral sensitivity range from 8 to 14 μm and provided with a germanium optical lens) for visible and thermal images. Auteurs used Gabor wavelet transform for feature extraction and Support Vector Machine (SVM) for classification. Experimental results achieve a recognition rate of 96, 4%.

Y.Jin et al. [13] have developed a Coupled Discriminative Feature Learning for Heterogeneous Face Recognition. They implement a method that represents the discriminative features of the face by building an optimal filter eigenvector with the raw pixels of the image. The developed approach uses Local Ternary Patterns for encoding local patterns and cosine metrics to estimate the similarities between images. The performance of this method was tested on CASIA 2.0 NIR-VIS database.

This database is widely used in publications. CASIA NIR-VIS 2.0 was collected from 2007 to 2010 by Stan Z. Li et al.[5], it contains visible and near infrared frontal images of 735 subjects. Two cameras were used to acquire the visible and near infrared images. The visual color face images are captured using Canon A640 Camera and home-brew device were used for near infrared image acquisition. The NIR imaging device used to capture NIR image is a standard version for indoor use and under fluorescent lighting. A long pass optical filter is integrated in this camera and allows capturing images in wavelengths 720, 800, 850 and 880 nm. The spatial resolution of acquired images is 640×480 images. Experimental results indicated that the accurate recognition rate is less than 90%.

In 2021 R. He et al. [14] have proposed a Coupled Adversarial Learning (CAL) system for semi-supervised heterogeneous face recognition. This approach has used the VIS-NIR face matching by performing adversarial learning on both image and feature levels. VIS images had been generated from the unmatched VIS-NIR images. This system did not acquire images in the capture module but used images from CASIA NIR-VIS 2.0 database for the other modules. A series of end-to-end neural network composed with 29 convolution layers with residual blocks (LightCNN-29) have been used to extract and learn features. The experimental results indicated a rank 1 accuracy from 98.6% to 99.6%.

A. Yu et al [15] have implemented face recognition system that used Generative Adversarial Networks(GANs) in the VIS and NIR. The VIS images have been generated from the NIR images. In order to reduce the domain gap between the NIR and VIS data, an attention module has been developed by the authors. This system has acquired images for visible and NIR range with two camera. A Large-Scale Multi-pose High-Quality Database of NIR-VIS images called LAMP-HQ was collected. A LightCNN-29 has been used as classifier. The performance achieved with this system has showed a rank 1 accuracy from 94.94% to 97.91%.

Song et al. [16] have implemented an adversarial discriminative feature learning framework for VIS and NIR face recognition. In order to compensate the detection gap the authors have applied the methods based on adversarial learning on both the raw pixel space and the compact feature space. This approach combines ResNet and Light CNN. The experiment has been performed on three databases: CASIA NIR-VIS 2.0 database, BUAA-VisNir [5] face database and Oulu-CASIA NIR-VIS facial expression database. Experimental results achieved give respectively for the tree databases a rank 1 accuracy of 98,15%, 95.2% and 95.5%. The BUAA-VisNir face database has been created by D. Huang et al. It contains 162 images/person of 150 persons. The images were acquired in visible and NIR range with 9 different facial expressions.

Oulu-CASIA NIR&VIS facial expressions database was set up by Chinese Academy of Sciences. It contains videos with six typical expressions i.e. happiness, sadness, surprise, anger, fear and disgust from 80 subjects captured with two imaging systems (SN9C201 & 202) which combines a NIR camera and a visible camera. This imaging system captures NIR and visible images under three different illumination conditions

normal indoor illumination, weak illumination and dark illumination. The imaging hardware works at the rate of 25 frames per second and the image resolution is 320×240 pixels.

In order to correct misalignment problems between visible and NIR matched images P. Zao et al. [17] have developed a Self-Aligned Dual NIR-VIS Generation for Heterogeneous Face Recognition. The architecture proposed by the authors is based on GANS and allows generating semantically aligned dual NIR-VIS images with the same identity. This system has not acquired images in the capture module but has used images from CASIA NIR-VIS 2.0, Oulu-CASIA NIR-VIS and BUAA VIS-NIR. The features have been extracted with lighCNN and two encoder networks for generation tasks. A rank 1 accuracy close to 99.9% has been performed for each datasets.

M. Diarra et al. [18] have proposed the MS-FRHF (Multispectral Face Recognition using Hybrid Feature) approach for visible and thermal. They used Robotics Intelligent System (IRIS) database for feature extraction. The points of interest and the texture have been extracted respectively with the Maximally Stable Extremal Region (MSER) keys points extractor and Gray Level Co-Occurrence Matrix (GLCM). Principal Component analysis (PCA) was used to fuse the feature. Authors have concluded this approach gives the best recognition rates than those obtained in the visible and thermal infrared.

In 2020, Zhihua Xie et al. [19] have developed the fusion methods based on the local binary model and the discrete cosine transform for face recognition in the infrared and visible range. For this purpose, the low frequency information is first extracted from the near-infrared images with the discrete cosine transform and the LBP is applied to represent the discriminative features. Then the features of the visible images have been extracted with the LBP and finally a fusion has been done. Experimental results have shown that the recognition rate has been improved with this approach.

Guo et al. [20] have proposed Face recognition system using both visible light image and near-infrared image and a deep network. This system uses two different cameras, one in the visible and the other in the near infrared. The visible and near infrared features have been first extracted with the neural network. Then the authors have used the cosine distance to determine the classification score. Finally, a fusion of the classification scores has been performed. They have used HIT LAB2 and SunWin Face database to test the performance of their model. Accuracy of 99.89% and 99.56% has been achieved on the two databases respectively in weak light change.

Hu et al. [21] have presented a Discriminant Deep Feature Learning based on joint supervision Loss and Multi-layer Feature Fusion for heterogeneous face recognition. This approach implements Convolutional Neural Networks by integrating a loss function called scatter loss in order to improve the discriminating power of the learned features in depth. The features extracted by the CNNs in the different visible and near infrared bands have then been merged. The performance of the system has been tested on CASIA NIR-VIS 2.0 and Oulu-CASIA NIR-VI databases. The experimental

results have given a rank 1 accuracy of 98.5% to 98.8% on the CASIA NIR-VIS 2.0 dataset, and of 98.5% to 99.3% on Oulu-CASIA NIR-VIS database.

F. Wei et al. [22] have developed an intraspectrum discrimination and interspectrum correlation analysis deep network (IDICN) approach for facial recognition. This system has improved the performance of multispectral face recognition by including inter- and intra-spectral information. Authors didn't acquire images but have used three databases: Hong Kong Polytechnic University (HK PolyU) dataset, Carnegie Mellon University (CMU) dataset and the University of Australia (UWA) dataset. This approach consists of a set of spectrum-set-specific deep convolutional neural networks with a spectrum pooling layer. The convolutional neural networks extract features related to a set of spectra, and the spectrum pooling layer selects a group of spectra with discriminative capabilities.

The HK PolyU dataset consists of 48-subject hyper-spectral image cubes, which are acquired using CRIs VariSpec liquid crystal tunable filter (LCTF) under halogen light. The spectral range extends from 400 to 720 nm with a step size of 10 nm.

CMU database is collected with a prototype spectropolarimetric camera developed by CMU. It contains the images of 54 subjects. The hyper-spectral range is between 450 and 1090 nm with a step length of 10 nm.

The UWA dataset consists of 120 hyper-spectral image cubes of 70 subjects acquired with the VariSpec LCTF CRIs integrated with a photonic focusing camera. Each hyperspectral image cube contains 33 bands covering the spectral range from 400 to 720 nm with a 10 nm step size.

Experimental results have achieved average recognition rates of 99.76%, 100% and 99.85 respectively on the three bases.

III. MATERIALS AND METHODS

A. Our One-shot Multispectral Filter Array Acquisition System

This section describes the Multispectral Filter Array one shot camera used for image database collection.

In recent years, sustained research efforts have been carried in the field of multispectral imaging systems incorporating MSFA. Multi-spectral imaging using a single camera with MSFA is an efficient way to acquire spectral data. It has the potential to promote a fast and real time multispectral imaging system. The concept of Spectral (or Multispectral) Filter Arrays has been developed recently and enables one shot multispectral acquisition with a compact camera design. Generally, Multi-Spectral filter array (MSFA) is made up recurrent patterns of filtering elements. Multi-Spectral filter array (MSFA) cameras are a new single-shot spectral imaging technology that is defined by a basic repetitive pattern composed of filter elements. Each filtering element is sensitive to a specific spectral band. Multi-spectral Filters Array are filter matrix in which each filter corresponds to a spectral band. During the technical design of the cameras the filters are carefully selected. MSFA one-shot camera architecture shows

that the MSFA overlap the camera sensor so as to cover it. The light entering the camera is filtered with a band pass spectral filters on each pixel. A MSFA aims at object property estimation and/or objective color measurement. A MSFA might be defined by its moxel, mosaic element, which corresponds to the occurrence of a pre-defined pattern that consist of a set of filters arranged geometrically in a relative manner [23]. The moxels or multispectral pixels are the smallest patterns in the MSFA. An overview of the global approach is shown in following figure (Fig. 2).

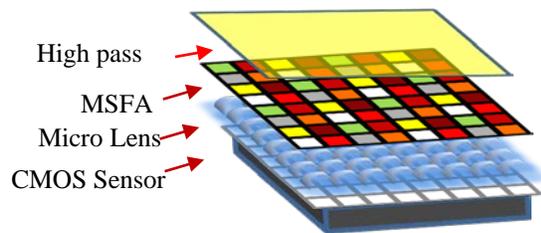


Fig. 2. Global Scheme of MS Imaging System with MSFA.

Multispectral Filter Array Camera was used to create the multispectral images database. The oneshot acquisition system has been designed in the Electronics, Computer and Image (LE2I) Laboratory which is now ImViA (Imaging and Artificial Vision) during UE H2020 project called EXIST (EXtended Image Sensing Technologies). It is a compact and lightweight acquisition system that integrates a single Viimagic 9220H sensor, MSFA for one-shot imaging system, Optic lenses, Electronic board for driving the sensor and Camera board for image acquisition. In order to correct the linearity of the sensor and also to measure the spectral sensitivity, a characterization of the CMOS sensor has been made before mounting the Multi-Spectral filter Arrays. The MSFA has been selected carefully considering a regular distribution of pixels in the moxel. Our personalized filter matrix was built using SILIOS technologies. SILIOS Technologies has developed the COLOR SHADES® technology, which use Fabry-Perot interferometer to manufacture multispectral transmittance filters. This technology is based on the combination of thin film deposition and micro- / nano-etching processes on a fused silica substrate. Standard micro-photolithography steps are used to define the cell geometry of the multispectral filter. COLOR SHADES® provides band pass filters originally in the visible range from 400 nm to 700 nm. SILIOS has developed filters in the NIR range in collaboration with LE2I (ImViA) laboratory, combining their technology with a classical thin film interference technology to realize our filters. The MSFA system, integrated into a camera with dedicated hardware and software computations, allows operating in real-time application with 30 fps. The filters used overcome the problems caused by lighting variation, motion blur noise and SNR noise which severely affect the performance of facial recognition systems using CMOS. These are 8 optimal filters selected in the wavelengths {685, 720, 770, 810, 835, 870, 895, 930} (in nm) thanks to a technical study carried out at LE2I laboratory. Fig. 3 illustrates the spatial distribution of moxel.

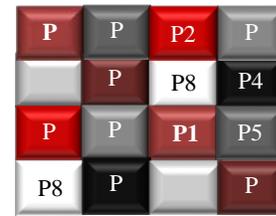


Fig. 3. Final Moxel of MSFA.

This acquisition system is a light robust, broadband multispectral system which tends to measure the physical properties of the object. It is also a real time system which covers visible and near infrared spectra (650 to 950 nm). The MSFA integrated in our acquisition system has a small moxel 4x4 with 2 pixels per band and the size of filter pitch is 5x5 μm^2 . At each acquisition, the MSFA one shot camera provides a best resolution of raw or mosaic image of size 2072 x 1104 pixels in which the values of every channel are accessible at each pixel according to the MSFA. The missing channel values are there after estimated by demosaicking process. In order to provide an optimal solution for the loss of spatial resolution inherent to MSFA, specific algorithms have been developed for multispectral demosaicking. Indeed, the demosaicking process must be associated with the design of the MSFAs otherwise the loss of image resolution could be critical. As the acquisition system is an MSFA one shot camera, it privileges spectral resolution. The Fig. 4 presents the MSFA camera.



Fig. 4. MSFA One-Shot Camera.

B. Data Collection

Multispectral face images were collected over two years in Imaging and Artificial Vision (ImViA) Laboratory in Faculty of Science and Technology of Burgundy University in France. The acquisition room is a black room, dedicated to MS imaging. The photos were taken with an illuminant light with different orientations. This light illuminates the subject's face to be photographed. The light is oriented from left to right and vice versa during the acquisition. The distance between the camera and the subject to be acquired is 1 meter. The relative position of camera is shown in Fig. 5.

The MS images database have been acquired in winter 2020, winter 2021 and spring 2021. Participants were made up of residents and international students, and 75% of subjects agreed to have their photo posted. They are men and women of all ages, black and white. Participants are from different continents namely Europe, Asia, Africa, Arabic and African. The multi-spectral image database named EXIST MS database contains faces images of 103 subjects. Face images MS database is structured as follows:

- All images are tiff format and size 2072 x 1104;

- Our database contains 20 different mosaic faces images per subject;
- 20x8 demosaic faces images per subject.

In total our MS images database contains 103x20x8 (16 480) MS images.

The following figure (Fig. 6) shows some mosaic face images of a subject in the database.



Fig. 5. Acquisition Set-up in the Black Room.

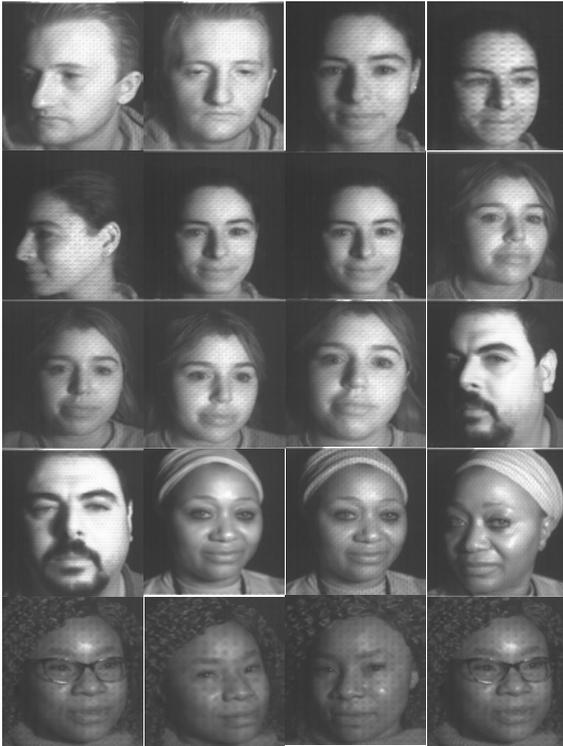


Fig. 6. Sample Images from our MS Database.

C. Image Demosaicking Algorithm

Having used a multispectral acquisition system which provides a mosaic image, a demosaicking is necessary to

generate multi-band images. In fact, a mosaic or raw image is of size $(X \times Y)$ pixels, in which a single band $k \in \{1, \dots, K\}$ refers to the value of each pixel p depending on the MSFA structure. Image demosaicking is a process that separates mosaic or raw images into multispectral images according to filter number in the MSFA. Before the demosaicking, a strip extraction was done. It consists of multiplying the mosaic image by different binary masks $M^k(x,y)$, $k \in \{1,2, \dots, K\}$ [24](pp. 31-34).

These masks have the value 1 at the positions where the component is available, and 0 at the other positions. Each plane component is obtained after multiplying the mosaic image by the corresponding mask M^k .

$$M^k_{(x,y)} = \begin{cases} 1 & \text{si } (x \bmod \sqrt{K}) + (y \bmod \sqrt{K}) \times \sqrt{K} = k \\ 0 & \text{sinon} \end{cases} \quad (1)$$

In this case, the multiplication of the mosaic image by each mask allows us to obtain 8 planes of shifted images in which only one component is available at each pixel. Each mask corresponds to an image plane I'^k .

$$I'^k = I \odot M^k \quad (2)$$

Where \odot denotes the element-wise product and M^k is a binary mask defined at each pixel p .

Bilinear interpolation method is used for multispectral demosaicking. This method enables to estimate the missing in each pixel. Bilinear interpolation can be expressed as a resampling technique based on distance weighted average of the four nearest pixel values to evaluate each missing pixel value. Bilinear interpolation is a succession of two linear interpolations, each in one direction. The linear interpolations can be performed in multiple directions. For a missing pixel $P(i,j)$ at position (i,j) , the linear interpolation is defined as follows:

- Diagonally

$$P(i,j) = \frac{1}{4} \sum_{(m,n) \in \{(-1,-1), (-1,1), (1,-1), (1,1)\}} p(i+m, j+n) \quad (3)$$

- Vertically

$$P(i,j) = \frac{1}{2} \sum_{(m,n) \in \{(-1,0), (1,0)\}} p(i+m, j+n), \quad (4)$$

- Horizontally

$$P(i,j) = \frac{1}{2} \sum_{(m,n) \in \{(0,-1), (0,1)\}} p(i+m, j+n) \quad (5)$$

The multispectral image demosaicking using bilinear interpolation also consists in convolving each component plane obtained by an H filter. This filter is determined as a function of the spatial distance between the neighbors from the central pixel.

Two filters H1 and H2 (Fig. 7) were used to do the convolution. The interpolated image band is defined by

$$I^k = I'^k \odot H \quad (6)$$

With $H=H1$ or $H=H2$

1/9	2/9	1/3	2/9	1/9
2/9	4/9	2/3	4/9	2/9
1/3	2/3	1	2/3	1/3
2/9	4/9	2/3	4/9	2/9
1/9	2/9	1/3	2/9	1/9

1/5	1/4	1/5	1/3	0
0	1/3	1/5	1/4	1/5
1/5	1/3	1	1/3	1/5
1/5	1/4	1/5	1/3	0
0	1/3	1/5	1/4	1/5

Fig. 7. Filter H1 and H2.

To estimate the missing pixel value $P(i,j)$, the image I^k is convolved with each of the H1 and H2 filters. If V1 and V2 are convolution results then the missing pixel value is defined by average of V1 and V2. This process is used to estimate each missing pixel in I^k . The figure (Fig. 8) shows the missing pixel value estimation in demosaicing process.

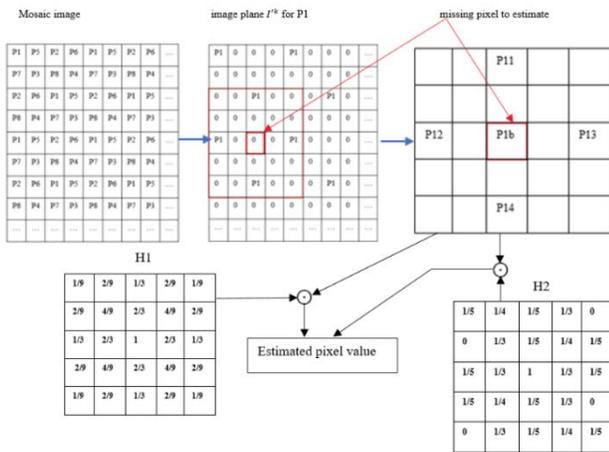


Fig. 8. Missing Pixel Estimation for Demosaicing Process.

The image demosaicing generates 8 images belonging to the wavelengths {685, 720,770, 810, 835, 870, 895, and 930} (in nm). The figure (Fig. 9) illustrates the demosaicking images process.

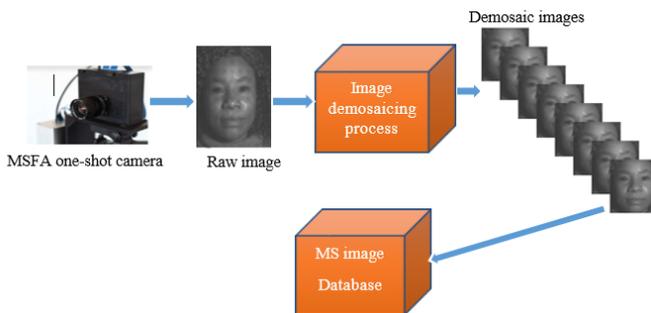


Fig. 9. The Process of Multispectral Images Demosaicking.

D. Methodology

This section describes the algorithms used for facial recognition with the Multi-Spectral Filters Array camera described above. Fast Discrete Curvelet Transform (FDCT) and Convolutional Neural Networks are used to implement the facial recognition system. In order to exploit the important information contained in each spectral band we first use a

fusion at the image level with the FDCT. First, the image-level fusion method is implemented with FDCT. Then, VGG19 and ResNet 101 neural networks are used to perform the recognition. The following figure (Fig. 10) illustrates the proposed method:

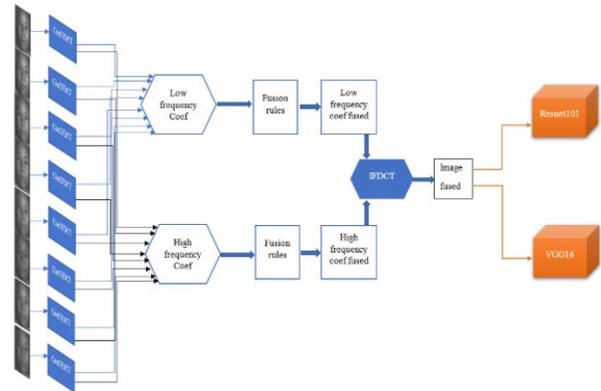


Fig. 10. Bloc Diagram of the Proposed Face Recognition Method using MSFA One-shot Camera.

The Fast Discrete Curvelet Transform (FDCT) is a wavelet transform that decomposes an image in low and high frequency. FDCT adopts local Fourier transform for the frequency domain decomposition. First, each of the eight demosaic image was decomposed with the FDCT. Then the low frequency coefficients are merged together and so for the high frequency coefficients. Inverse Fast Discrete Curvelet Transform was applied to obtain a merged image containing the information of all bands. Finally transfer learning for VGG19 and ResNet 101 are used to classify the images fused with FDCT method.

IV. EXPERIMENTS AND RESULTS

EXIST is the image database acquired with the described acquisition system and the one used evaluation purpose.

All the images are acquired from 20 different positions per person; in total 2000 images in the multi-spectral images database are taken.

Experimentations are carried out on Microsoft System windows, version 2010, with two computers. The first one was equipped with an Intel (R) Core (TM) i7-8565U CPU, 8 GB of RAM memory. The second has a graphical processing unit (GPU) NVIDIA Quadro P400 with 32GB of Random Access Memory (RAM). All the code are developed in the programming language of Matlab 2020 and Python 3.7.

Table I describes the training parameters for VGG19 and ResNet101.

To analyse the results, the following performance evaluation metrics have been calculated: accuracy, precision, recall, F1 score, Matthews Correlation Coefficient (MCC) and Means Square Error (MSE).

An accuracy indicates the percentage of correct predictions.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

TABLE I. PARAMETERS USED IN THE TRAINING PROCEDURE

Parameters	CNN	
	VGG19	ResNet101
Batch size	32	16
Optimization algorithm	SGD ^a	ADAM
Learning rate	0.0001	0.0001
Epoch number	20	20

^a Stochastic Gradient Descent

Where TP (True Positive), TN (True Negative), FP (False Positive), FN (False Negative).

The precision is the proportion of true positives out of all detected positives.

$$precision = \frac{TP}{TP+FP} \quad (8)$$

The recall is the number of true positives that are correctly classified.

$$recall = \frac{TP}{TP+FN} \quad (9)$$

Component F1 score includes recall and precision and is calculated as

$$F1_{score} = \frac{2*precision*recall}{precision+recall} \quad (10)$$

The Matthews Correlation Coefficient (MCC) is the method of calculating correlation coefficient between real and predicted values. MCC is more informative score and give best result in binary classification assessment [25]. The range of values of MCC is between -1 and 1. A model with a score of 1 is a perfect model and -1 is a poor model.

$$MCC = \frac{(TP*TN)-(FP*FN)}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \quad (11)$$

The Mean Squared Error (MSE) allows to calculate error between predict values \hat{y} and reals value y .

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (12)$$

Where N is the number of samples we are testing against.

The different metrics calculated allow for a better interpretation of the results. The accuracy of a model predicts the percentage of all persons that can be recognized by the model. The recall indicates the number of correct predictions that were actually recognized.

The fusion rules such as min, max and average are used to obtain the images. Three types of experiments were done: a first one with the images obtained with FDCT and the min fusion rule, a second one with the images obtained with FDCT and the max fusion rule and a third one with the average fusion rule.

In Table II and Table III, the performance of the VGG19 and ResNet101 model for min, max and average fusion are listed respectively.

TABLE II. VGG19 RESULTS

Fusion	Metrics				
	Precision	MSE	F-score	Recall	MCC
average	1.0	2.80e-07	1.0	1.0	1.0
min	0.9849	7.96e-07	0.98	0.98	0.98
max	0.9870	6.76e-07	0.98	0.98	0.98

TABLE III. RESNET101 RESULTS

Database	Metrics				
	Precision	MSE	F-score	Recall	MCC
average	0.99	4.91e-10	0.99	0.99	0.99
min	0.97	7.98e-10	0.97	0.97	0.97
max	0,96	91e-10	0,96	0,96	0,96

Note that the two models VGG19 and ResNet101 have been trained with the images of the databases, taking into account different batches size 8, 16 and 32.

By comparing the different metrics calculated on the two neural networks, the results show that the performance of average fusion rule method are superior to those of min and max fusion rule.

Table IV describes the recognition rate of stat of art recognition and the proposed system.

TABLE IV. THE RECOGNITION RATES

Acquisition system	Recognition methods	Recognition Rate
Visible camera + NIR camera + thermal camera	Gabor wavelet transform and Support Vector Machine (SVM) [12]	96.4%
Visible camera and NIR camera	Local Ternary Patterns and cosine metrics [13]	90%
Visible camera and NIR camera	LightCNN-29 [14]	98.6% to 99.6% ^a
Visible camera and NIR camera	LightCNN-29 [15]	94.94% to 97.91%.
Visible camera and NIR camera	ResNet and LightCNN-29[16]	98.15%, 95.2% and 95.5% ^b
Visible camera and NIR camera	lighCNN and two encoder networks[17]	99.9%
Visible camera and NIR camera	Neural Network and cosine distance[20]	99.89% and 99.56%
Visible camera and NIR camera	CNN[21]	98.5% to 98.8% and 98.5% to 99.3%
Hyper-spectral camera	deep convolutional neural networks[22]	99.76%, 100% and 99.85
EXIST camera	FDCT and VGG19	100%
EXIST camera	FDCT and ResNet 101	99%

^a Results achieved between two values

^b Results achieved on different databases

Table IV indicates that most facial recognition systems in the visible and NIR range use multiple cameras. Depending to the image database used, the recognition rate range 95.3% to 100%. The rate of 100% was reached with the deep convolutional neural network algorithm. In general systems using neural networks algorithm have a rate close to 100%. In this case, depending to the recognition algorithm, the rate is range 90% to 100%. The face recognition systems that use an acquisition system integrating Multispectral Filters Arrays (MSFA) achieve perform as well as those using several cameras.

V. DISCUSSION

In this article, face recognition based on MSFA oneshot camera were demonstrated. Most previous studies in the literature used multiple cameras and Convolutional neural to extract features for face identification. Y.Jin et al. in [13] used multiple cameras with Local Ternary patterns, cosine metrics and achieve recognition rate of 90%. The recognition systems presented in the literature that are based on Gabor wavelet Transform and Support Machine Vector (SVM) achieve a recognition rate of 96.4% [12]. The recognition systems based on Convolutional Neural Network [14],[15],[16],[18],[22] achieve respectively accuracy in [98.6% - 99.6%], [94.94% - 97.91%] and [98.15%, 95.2%,95.5%], [99.89%, 99.56%],[98.5%-98.8%, 98.5%-99.3%] depending on database. Also [23] used hyperspectral camera with neural networks and get [99.76%, 100%, 99.6%] accuracies on three different databases. Results and experiments of the facial recognition using MSFA oneshot camera achieve accuracies of 99% and 100% with respectively Resnet101 and VGG19.

The comparison of the results shows that the different facial recognition algorithms implemented give good performance depending on the images and methods used. The proposed system reaches one of the best performances. But also because of its camera and its algorithms, it is the only system which proposes a real time acquisition of images.

VI. CONCLUSION

This paper presents a new multispectral facial recognition system using one shot multispectral imaging systems integrated Multispectral Filters Array for acquisition. It is a one-shot acquisition system that operates in real time on the visible and NIR spectra. A multispectral database containing images with spectral information has been created for this end. This recognition system is based on the Fast Discrete Curvelet Transform, ResNet 101 and VGG19 Convolutional Neural Network. Experimental results show that face recognition systems using MSFA cameras perform as well as those using multiple cameras. This system is however more interesting as it is more economical, technically reliable and especially equipped with a real time acquisition system.

In future work, the image database will be extended; other multispectral demosaicing and recognition algorithms will be implemented.

REFERENCES

- [1] L. Kambi Beli and C. Guo, "Enhancing Face Identification Using Local Binary Patterns and K-Nearest Neighbors," *Journal of Imaging*, vol. 3, no. 3, Art. no. 3, Sep. 2017, doi: 10.3390/jimaging3030037.
- [2] M. Lal, K. Kumar, R. H. Arain, A. Maitlo, S. A. Ruk, and H. Shaikh, "Study of Face Recognition Techniques: A Survey," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 6, Art. no. 6, 29 2018, doi: 10.14569/IJACSA.2018.090606.
- [3] R. He, X. Wu, Z. Sun, and T. Tan, "Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1761–1773, Jul. 2019, doi: 10.1109/TPAMI.2018.2842770.
- [4] Z. Xie, P. Jiang, and S. Zhang, "Fusion of LBP and HOG using multiple kernel learning for infrared face recognition," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, May 2017, pp. 81–84. doi: 10.1109/ICIS.2017.7959973.
- [5] L. L. Chambino, J. S. Silva, and A. Bernardino, "Multispectral Facial Recognition: A Review," *IEEE Access*, vol. 8, pp. 207871–207883, 2020, doi: 10.1109/ACCESS.2020.3037451.
- [6] M. Mateen, J. Wen, Nasrullah, and M. A. Akbar, "The Role of Hyperspectral Imaging: A Literature Review," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 8, Art. no. 8, 49/01 2018, doi: 10.14569/IJACSA.2018.090808.
- [7] Y. Yang, S. Tong, S. Huang, P. Lin, and Y. Fang, "A Hybrid Method for Multi-Focus Image Fusion Based on Fast Discrete Curvelet Transform," *IEEE Access*, vol. 5, pp. 14898–14913, 2017, doi: 10.1109/ACCESS.2017.2698217.
- [8] T. Gwyn, K. Roy, and M. Atay, "Face Recognition Using Popular Deep Net Architectures: A Brief Comparative Study," *Future Internet*, vol. 13, no. 7, Art. no. 7, Jul. 2021, doi: 10.3390/fi13070164.
- [9] H. Ling, J. Wu, L. Wu, J. Huang, J. Chen, and P. Li, "Self Residual Attention Network for Deep Face Recognition," *IEEE Access*, vol. 7, pp. 55159–55168, 2019, doi: 10.1109/ACCESS.2019.2913205.
- [10] Y. Park and B. Jeon, "An Acquisition Method for Visible and Near Infrared Images from Single CMYK Color Filter Array-Based Sensor," *Sensors*, vol. 20, no. 19, p. 5578, Sep. 2020, doi: 10.3390/s20195578.
- [11] X. Chen, H. Wang, Y. Liang, Y. Meng, and S. Wang, "A Novel Infrared and Visible Image Fusion Approach Based on Adversarial Neural Network," *Sensors*, vol. 22, no. 1, Art. no. 1, Jan. 2022, doi: 10.3390/s22010304.
- [12] Z. Abood, G. Karam, and R. Haleot, "Face Recognition Using Fusion of Multispectral Imaging," 2017, p. 112. doi: 10.1109/AIC-MITCSA.2017.8722957.
- [13] Y. Jin, J. Lu, and Q. Ruan, "Coupled Discriminative Feature Learning for Heterogeneous Face Recognition," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 3, pp. 640–652, Mar. 2015, doi: 10.1109/TIFS.2015.2390414.
- [14] R. He, Y. Li, X. Wu, L. Song, Z. Chai, and X. Wei, "Coupled adversarial learning for semi-supervised heterogeneous face recognition," *Pattern Recognition*, vol. 110, p. 107618, Feb. 2021, doi: 10.1016/j.patcog.2020.107618.
- [15] A. Yu, H. Wu, H. Huang, Z. Lei, and R. He, "LAMP-HQ: A Large-Scale Multi-pose High-Quality Database and Benchmark for NIR-VIS Face Recognition," *Int J Comput Vis*, vol. 129, no. 5, pp. 1467–1483, May 2021, doi: 10.1007/s11263-021-01432-4.
- [16] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial Discriminative Heterogeneous Face Recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Art. no. 1, Apr. 2018.
- [17] P. Zhao, F. Zhang, J. Wei, Y. Zhou, and X. Wei, "SADG: Self-Aligned Dual NIR-VIS Generation for Heterogeneous Face Recognition," *Applied Sciences*, vol. 11, no. 3, Art. no. 3, Jan. 2021, doi: 10.3390/app11030987.
- [18] M. Diarra, P. Gouton, and A. K. Jérôme, "Multispectral face recognition using hybrid feature," *Electronic Imaging*, vol. 2017, no. 18, pp. 200–203, Jan. 2017, doi: 10.2352/ISSN.2470-1173.2017.18.COLOR-061.
- [19] Z. Xie, L. Shi, and Y. Li, "Two-Stage Fusion of Local Binary Pattern and Discrete Cosine Transform for Infrared and Visible Face Recognition," in *Emerging Trends in Intelligent and Interactive Systems and Applications*, Cham, 2021, pp. 967–975. doi: 10.1007/978-3-030-63784-2_117.
- [20] K. Guo, S. Wu, and Y. Xu, "Face recognition using both visible light image and near-infrared image and a deep network," *CAAI Transactions*

- on Intelligence Technology, vol. 2, no. 1, pp. 39–47, Mar. 2017, doi: 10.1016/j.trit.2017.03.001.
- [21] W. Hu and H. Hu, “Discriminant Deep Feature Learning based on joint supervision Loss and Multi-layer Feature Fusion for heterogeneous face recognition,” *Computer Vision and Image Understanding*, vol. 184, pp. 9–21, Jul. 2019, doi: 10.1016/j.cviu.2019.04.003.
- [22] F. Wu et al., “Intraspectrum Discrimination and Interspectrum Correlation Analysis Deep Network for Multispectral Face Recognition,” *IEEE Transactions on Cybernetics*, vol. 50, no. 3, pp. 1009–1022, Mar. 2020, doi: 10.1109/TCYB.2018.2876591.
- [23] P.-J. Lapray, X. Wang, J.-B. Thomas, and P. Gouton, “Multispectral Filter Arrays: Recent Advances and Practical Implementation,” *Sensors*, vol. 14, no. 11, pp. 21626–21659, Nov. 2014, doi: 10.3390/s141121626.
- [24] S. Mihoubi, “Snapshot multispectral image demosaicing and classification,” *Theses, Université de Lille*, 2018.
- [25] D. Chicco, M. J. Warrens, and G. Jurman, “The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen’s Kappa and Brier Score in Binary Classification Assessment,” *IEEE Access*, vol. 9, pp. 78368–78381, 2021, doi: 10.1109/ACCESS.2021.3084050.

Detecting Distributed Denial of Service in Network Traffic with Deep Learning

Muhammad Rusyaidi¹, Sardar Jaf²

Faculty of Technology, School of Computer Science
Sunderland University, UK, Sunderland
SR6 0DD, United Kingdom

Zunaidi Ibrahim³

Mechanical Engineering, Universiti Teknologi Brunei
Tungku Highway, Gadong BE1410
Brunei Darussalam

Abstract—COVID-19 has altered the way businesses throughout the world perceive cyber security. It resulted in a series of unique cyber-crime-related conditions that impacted society and business. Distributed Denial of Service (DDoS) has dramatically increased in recent year. Automated detection of this type of attack is essential to protect business assets. In this research, we demonstrate the use of different deep learning algorithms to accurately detect DDoS attacks. We show the effectiveness of Long Short-Term Memory (LSTM) algorithms to detect DDoS attacks in computer networks with high accuracy. The LSTM algorithms have been trained and tested on the widely used NSL-KDD dataset. We empirically demonstrate our proposed model achieving high accuracy (~97.37%). We also show the effectiveness of our model in detecting 22 different types of attacks.

Keywords—Cybersecurity; Cyber-attack; DDoS attack; machine learning; deep learning; recurrent neural networks; long short-term memory

I. INTRODUCTION

COVID-19 pandemic has caused a great deal of fear, worry, and a significant shift in our way of life. Organizations have had to adapt to the requirement for remote working on a large scale and rapidly. Because COVID19 has produced or expanded applications and use cases of digital technologies, this pandemic is proven to be a motivator for digital transformation. Despite the pandemic, the world can still interconnect with each other through a network with rapid development in IoT4.0 technologies. This involves millions of data bytes being produced, processed, converted, exchanged, or shared and utilized to produce an outcome in specific applications. This involves the security elements to protect sensitive data and the privacy of each individual user of cyberspace or network. Distributed Denial of service (DDoS) is a type of attack in which the victim's resources are depleted, rendering them unable to handle valid requests. Nonetheless, the number of DDoS attacks and the amount of DDoS traffic are increasing, requiring more research into the detection of such security risks. Therefore, the use of machine learning to ensure the intensity of this data is very important. In this study, we examine network traffic behaviors for cyber detection by the application of various machine learning algorithms to improve the accuracy of DDoS attack detection.

DDoS attacks are common network exploitation type of cyber-attack. The attacker creates network exhaustion to legitimate users by causing a computer or network system to

crash, stopping them from accessing server or the Internet, either temporarily or continuously. According to Singh et al. [1], the DDoS attack is one of the most common and major cyber-attacks. Ray et al. [2] also states that more advanced technology is needed to improve DDoS attack detection in computer networks. Since detecting DDoS attacks is a difficult task before any mitigation measures can be performed, cybersecurity fundamentals are required to design a system that can detect threats. DDoS attacks were initially detected by traffic engineers using rule-based approach. This strategy have fallen behind the dynamic and evolving nature of DDoS attacks. Academics and industry are researching the prospect of integrating machine learning into DDoS detection process because of their immense potential and success in various Computing domains. Threats can be recorded more rapidly and correctly with machine learning algorithms, such as Naïve Bayesian, K-Nearest Neighbor, Random Forest and Recurrent Neural Network.

In this study, we focus on exploring the effectiveness of deep learning algorithms to improve the accuracy of DDoS attack detection in order to better analyze network traffic activities for cyber threat detection.; Also, we aim to discover a feature selection strategy that, when combined with a machine learning system, can improve DDoS detection accuracy rates. Our selected deep learning algorithm is based on a Recurrent Neural Network (RNN) classifier to distinguish between normal and attack traffic.

The remaining of this paper is organized as follow: Section II describes the literature review; Section III describes our methodology. The results from our experiments are presented in Section IV and we discuss our finding in Section V. We compare our results with previous research in Section VI. In Sections VII and VIII we conclude the paper and outline our future work, respectively.

II. LITERATURE REVIEW

Recent research has demonstrated the effectiveness of machine learning application in detecting DDoS attacks. In this section, Sambangi et al. [3] developed a machine learning model to predict DDoS and botnet attacks by using machine learning algorithm with multiple linear regression. They used the most widely used CICIDS 2017 benchmark dataset with entire packet payloads in pcap format, which is extensively used in labeled network flows. They also demonstrated that their machine learning model could detect DDoS attacks using

the regression analysis technique. Yuan et al. [4] showed that Recurrent Neural Network surpasses Random Forest in terms of generalization. The effectiveness of deep learning, where reduced the error rate from 7.517% to 2.103% using an ISCX2012 dataset, compared to traditional machine learning methods. They experiment uses the ISCX2012 dataset, made available by the University of New Brunswick in 2012. Guerre- Manzanares et al. [5] proposed the concept of employing hybrid feature selection models to lower the size of the feature to achieve more accurate results. The dataset contained 115 features. To limit the number of features, the filter; wrapper; and hybrid models were used for choosing the potential feature. These features were then loaded into a K-Nearest Neighbor (KNN) and Random Forest model, both of which had a high accuracy of 99%.

Sabeel et al. [6] presented the idea of using two deep learning models (deep neural network and long short-term memory) for binary prediction of unknown Denial of Service (DoS) and DDoS attacks. The models were evaluated on the benchmark CICIDS2017 dataset. According to Sabeel et al. [6], the models fail to detect unknown threats accurately. However, after retraining the deep learning models by merging newly synthesized datasets with the old ones, the True Positive Rate (TPR) achieved was 99.8% and 99.9% for DNN and LSTM, respectively. Rusyaidi et al. [7] demonstrated deep learning effectiveness in DDoS detection. Elsayed et al. [8] demonstrate that combining RNN with an autoencoder allows input traffic to be classified into two categories: normal and malicious. Elsayed et al. [8] and Catak et al. [9] have used Deep learning to deal with a high degree of complex nonlinear interactions. They were making it a possible tool for identifying network attacks. By using 70% of the input data for training, Elsayed et al [8]. model showed the best results when compared to existing traditional machine learning techniques, resulting in 99% accuracy in their proposed method.

The experiment conducted by Gadze et al. [10], they used RNN and LSTM in the software-defined networking (SDN) controller to identify and mitigate DDoS attacks. It was gathering certain network parameters when operating in a normal and also when subjected to DDoS attack. The number of packets received and transferred at each switch, the packet count (number of packets per flow), the protocol type (TCP, UDP, or ICMP), the Source IP, and the Destination IP were some of the main features. They looked at three different possibilities. In the first case, 80% of the data was used for training, while 20% was used for testing. In the second instance, 70% of the data was used for training and 30% for testing. In the third situation, 60% of the data was used for training and 40% for testing. RNN and were used for detecting and mitigating DDoS attacks. The 70/30 (train-test ratio) split yields improved model accuracy compared to the 80/20 and 60/40 split ratios.

Ugwu et. al. [11] compared the results of traditional machine learning algorithms such as Naive Bayes (NB), Decision Tree (DT), and Support Vector Machine (SVM). The suggested LSTM and Singular Value Decomposition (SVD): deep learning algorithms demonstrate a significant improvement. Data pre-processing is performed on the network data, which includes data normalization and feature

conversion methods. The normalization method requires limiting network feature values to a narrow range of values and feature conversion method requires transforming non-numeric feature values to numeric. Kasim [12] used dimensional reduction features in the autoencoder (AE) model and Support Vector Machine (SVM) classifier to classify encoded data as DDoS or normal. AE-SVM successfully distinguishes between normal and DDoS attack traffic. The min-max method was used to normalize their data between 0 and 1, and the training vectors for the AE model were created. With the encoding process, the trained model delivered feature learning and feature reduction. The results showed that the AE-SVM method performed well in terms of low false-positive DDoS detection rates and fast anomaly detection.

Gormez et al. [13] demonstrate that by using traditional machine learning algorithms, ensemble, and deep feature extraction methods, Bayesian optimization is faster than traditional grid search optimization. However, it requires more computing resources than the train-test step. The scikit-library of Python is used to implement the experiment classification methods. Network traffic packet data was captured and converted into connection records. They used three types of features: basic features, time-based features, and connection-based features. Basic features are characteristics that can be easily derived from packet headers by counting specific packet properties for the connection. Before evaluating a model's performance, hyper-parameter optimization allows researchers to fine-tune its hyper-parameters. The Bayesian optimization process is used to create samples of hyper-parameter values to locate the optimums.

Hossain et al. [14] highlighted that one of the most important criteria in evaluating the performance of network attack detection systems is the availability of labeled dataset. According to their experimental data, the optimal hyper-parameter combinations were used for constructing their robust intrusion detection system. LSTM multiclass classification was used in the experiment, with 80% of the dataset used for training and 20% for testing. Hyper-parameter adjustment was also used to investigate the performance. The experiment results demonstrated that deep learning models (LSTM) have become emerging technology for network attack detection systems.

III. METHODOLOGY

A. Dataset Preparation and Pre-Processing

The NSL-KDD dataset from the University of New Brunswick Lab, which includes 125,973 network packets with 22 different types of attacks, as shown in Table I.

According to Tang et al. [15], one of the most up-to-date datasets for Intrusion Detection System (IDS) evaluation is the NSL-KDD dataset. There are 41 features in this dataset, divided into three categories: fundamental, content-based, and traffic-based features. DoS, probe, U2R, and R2L are the four types of attacks. We use the NSL-KDDTrain+ dataset train our DDoS detection system, while the NSL-KDDTest+ dataset is used to test it. As a result, the NSL-KDDTest+ dataset is a useful indicator of a model's zero-day attack resistance. To distinguish between genuine and malicious traffic, the use of

DoS as a basis is utilized. Table II summarizes the dataset's features. These features are not ordered on a scale. These attributes are further passing to the next phase for normalization.

A dataset may have missing values, irrelevant features, categorical data, or other flaws that prevent a machine learning algorithm from analyzing it. In some cases, standardization, data normalization, and other issues might prevail in some circumstances. The NSL-KDDTrain+ customized datasets have missing values, irrelevant features, and an issue with the categorical column. The following data cleaning and preparation processes were included in this project and will be discussed in the paragraph below. The selected dataset contains 4,898,431 data records.

There are a few rows/columns in the customized dataset that do not have a number (NaN) or have infinite values. In the NSL-KDD dataset, not all values are filled, and some have strings. Research made by Nimbalkar et al. [17] shows that the captured network traffic is unsuitable for machine learning models due to noise, which includes NaN and missing data. These settings must be fixed before any further operations can be performed. To address the problem of NaN values, a variety of approaches can be used, as stated in Nimbalkar et al. [17]. One method involves removing rows or columns with a particular number of NaN values, while the other approaches involve replacing a missing value with another value, such as the mean, median, mode, or other statistical measures of a column, a row, or a group of data. The selection must be made wisely based on the information available about the dataset. We are replacing the NaN values with mean and median at the features in this project since there are some features that have missing values. Some columns do not include the information needed to classify traffic as normal or malicious. As a result, constant columns are useless for any detection process. In the dataset, there is one attribute, num_outbound_cmds, which is always 0 for all rows in the training and test data. We remove this attribute because it could otherwise result in performance degradation and unnecessary complications. Therefore, for algorithms that demand numerous samples of one or more-time steps and features, we reshaped two-dimensional data where each row represents a sequence of three-dimensional array.

TABLE I. 22 DIFFERENT TYPES OF ATTACKS ALZHRANI ET AL. [16]

Attack Categories	Training Set Attack Names	Test Set Attack Names
DoS	Back, land, Neptune, pod, smurf, teardrop	Back, land, Neptune, pod, smurf, teardrop, (mailbomb), process table, udpstorm, apache2, worm
Probe	Ipsweep, nmap, portsweep, satan	Ipsweep, nmap, portsweep, satan, mscan, saint
U2R	Buffer overflow, load module, perl, rootkit	Buffer overflow, load module, perl, rootkit, sqlattack, xterm, pst
R2L	ftp-write, guess-passwd, imap, multihop, phd, spy, warezmaster	ftp-write, guess-passwd, imap, multihop, phf, spy, warezmaster, xlock, xsnoop, snmpguess, snmpgetattack, HTTP tunnel, send-mail, named, warez client

TABLE II. FEATURE OF NSL-KDD DATASET

1	Duration
2	Protocol_type
3	Service
4	Flag
5	Src_bytes
6	Dst_bytes
7	Land
8	Wrong_fragment
9	urgent
10	Hot
11	Num_failed_logins
12	Logged_in
13	Num_compromised
14	Root_shell
15	Su_attempted
16	Num_root
17	Num_file_creations
18	Num_shells
19	Num_access_files
20	Num_outbound_cmds
21	Is_host_login
22	Is_guest_login
23	Count
24	Srv_count
25	Error_rate
26	Srv_error_rate
27	Rerror_rate
28	Srv_rerror_rate
29	Same_srv_rate
30	Diff_srv_rate
31	Srv_diff_host_rate
32	Dst_host_count
33	Dst_host_srv_counts
34	Dst_host_same_srv_rate
35	Dst_host_diff_srv_rate
36	Dst_host_same_src_port_rate
37	Dst_host_srv_diff_host_rate
38	Dst_host_error_rate
39	Dst_host_srv_error_rate
40	Dst_host_rerror_rate
41	Dst_host_srv_rerror_rate

Each sequence has several time steps, each with one observation which is a feature. There are enough data records in the NSL-KDD dataset for training and testing. The availability of data records and the lack of redundant records, which can prevent false detection of DDoS attack, help in improved learning accuracy. After the dataset had been cleaned and pre-processed, we trained and tested LSTM and RNN algorithms. The dataset is split to train and test sets. These sets are necessary for training the estimator and subsequently evaluating the performance of the associated model. In this project, we use NSL-KDDTrain+ for training and NLS-KDDTest+ for testing. However, a common practice is to split for training and testing machine learning algorithms, as has been done by Rusyaidi et al. [7] and Gadze et al. [10].

Creating a model for classification or other similar tasks is at the basis of machine learning-based work. This is what the training phase accomplishes. The training dataset, produced before in the data split phase, is used to train a machine learning algorithm on a section of the whole dataset. An algorithm that has been trained produces a model that has learned from the data. There are a variety of classification estimators available. In this work, RNN and LSTM algorithms were used. These estimators were chosen for their ease of use, widespread use in the literature, and solid performance in related work by Yuan et al. [4], Elsayed et al. [8] and Gadze et al. [10].

B. Deep Learning Model Development

Traditional feedforward neural networks have the problem of assuming data to be unrelated. The feedback loops of the hidden units are the major difference between a RNN and a feedforward neural network. RNNs can process a sequence of inputs and save their state while processing the next sequence of inputs in deep learning. The essential information is stored in the node's memory and will be used for learning in future time steps as shown in Nazih et al. [18]. However, RNN has some issues remembering long-term memories as stated in Staudemeyer et al. [19]. Thus, it does not work well with long sequences. As a result, problems with RNNs such as vanishing gradient and short-term memory, can be solved using a type of RNN known as Long Short-Term Memory Networks (LSTM).

According to Laghrissi et al. [20], LSTM is a Recurrent Neural Network that can recall more context information than RNN and select what information is significant and not important by using distinct cell states. Different gates and a cell state are included in the LSTM. Althubiti et al. [21] explain that LSTM has a sigmoid function that produces numbers between 0 and 1. If the activation function's value is 0, the information is lost; if the value is 1, the information is saved. The input gate changes the state of the cell. The previous hidden state and the current input are sent into the input gate. The tan and sigmoid activation functions are included, as well as their multiplied values. The cell state is now computed by adding the output of the input gate point by point. Finally, the output gates determine the value of the next concealed state.

The LSTM algorithm overcomes the limitation of RNNs by learning long-term dependencies. Another distinction is that, whereas RNNs have only one neural network layer.

LSTM has four neural network layers that interact with each other. In this project, the input and embedding layers are used first, followed by one LSTM layer with dense layers as depicted in Fig. 1. Note that mean absolute error is used as loss function, Adam as optimizer function, Accuracy as performance metrics.

LSTM is particularly suited for data sequence applications because of its unique design. Fig. 1 shows the model looks up the embedding for each character, converts two-dimensional data into a three-dimensional array, executes the LSTM batch size, timestep, and LSTM units with the embedding as input, then applies the dense layer to generate result accuracy prediction results. Many previous research Laghrissi et al. [20], Althubiti et al. [21], Gadze et al. [10], and Sabeel et al. [6]

indicated that LSTM is an effective approach for learning long-term dependencies and efficiently representing the relationship between current occurrences and historical events. In this paper, we adopted LSTM for the design of our deep learning architecture.

C. Training and Testing Proposed LSTM – RNN Model

Feature selection is a key issue in machine learning projects. Guerra-Manzanares et al. [5] highlighted that one aspect of dimensionality reduction is feature selection. Not all the features in a dataset are equally essential for detecting the attack. In many cases, increasing the number of characteristics above a particular threshold has no discernible effect on classification performance. It simply adds to the complexity and delays in performance. Not only that, but it may also lead to overfitting and a decline in classification performance. As a result, we look for a minimal number of features that can appropriately identify the traffic in a dataset wherever possible. For this, we use the wrapper technique, namely correlation feature selection, which is supported in Scikit-learn.

The easiest strategy to train a model is to train the specific attack types to avoid being attacked by the same sort of attack. The 22 various forms of attacks (as shown in Table I) were utilized for training the model to reinforce it. The attributes of each attack have distinct values. These features and attack types were part of the training set, which was 80% of the full NSL-KDDTrain+ dataset. We use 20% of the NSL-KDDTest+ database for testing our model. The test set is separate from the training set.

D. Proposed Model Architecture

Fig. 2 shows the main components of the proposed system: the pre-processing, adaptive attribute selection, and classification of DDoS attack type. The procedure of subsystems is divided into three stages:

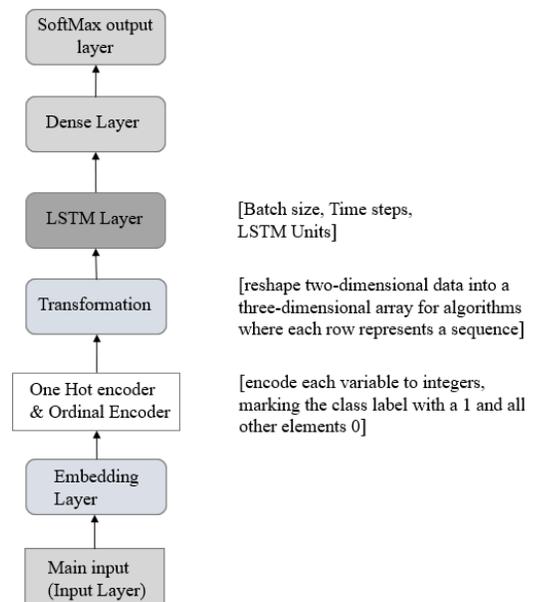


Fig. 1. The Proposed System Architecture representing each Layer.

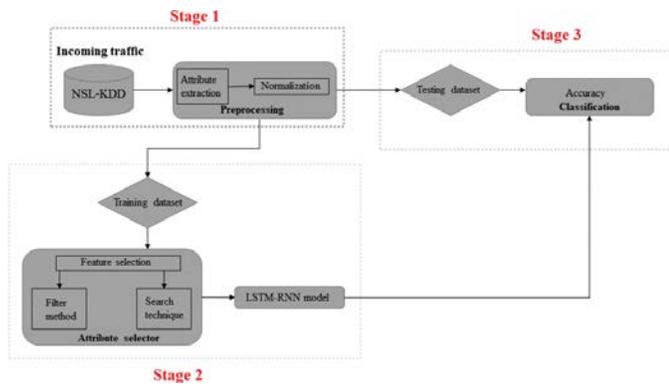


Fig. 2. The Main Components of the Proposed System Architecture with Representing the Processing Stage.

Stage 1: The Pre-processing stage involves collecting and normalizing attributes from network traffic. Data is separated into subgroups for training and testing. 80% of the data in NSL-KDDTrain+ is set as a training dataset for use in attribute selector (Stages 2), while the remaining 20% of the data in NSL-KDDTest+ is set to test dataset for use in Stage 3.

Stage 2: Various automatic threshold procedures are used in the Attribute Selection Subsystem to determine the minimum number of attributes.

Stage 3: classification and detection of DDoS attacks.

In order to have a more practical structure of the results, all of the experiments were organized into three stages, as shown in Fig. 2. Two experiments were carried out in Stage 1. These tests were conducted using estimators set to their default settings. No extra parameter tuning or feature selection work was done here; instead, a basic percent split technique was applied. A series of feature selection experiments were carried out in Stage 2. Once again, a percent split technique was applied without taking cross-validation into account. In Stage 2, multiple experiments comprising a feature selection operation were carried out.

Finally, the classification is in charge of detecting traffic data as DDoS in Stage 3. The results of stages 1, 2, and 3 were successfully achieved. The proposed machine learning model improved the DDOS attack detection approaches and increased the DDOS detection accuracy with a combination of features selection, adam optimizer, mean absolute error, oneHotEncoder strategy. Hence, there are test accuracy results after being implemented in the module. In the sections below, discussions are included that go along with it.

IV. EXPERIMENTAL RESULTS

We have implemented our deep learning architecture in TensorFlow with Keras backend. We used the mean absolute error approach to verify the model's loss while learning the deep neural network, which comprises two hidden layers. The "Adam" optimization function was used. "one-hot encoder" and "Ordinal Encoder" libraries from the "Sklearn" library also have been used to convert order-like values to numeric numbers. We trained the model with 150 epochs with a batch size of 44.

The training accuracy of the model used the sample loss and accuracy using a batch size of 44, as shown in Fig. 5. We have experimented with varying learning rates, but the selective learning rate of 0.013 produced the best result in our experiment. After 150 epochs, where the training and validation performance converged, the model achieved the highest training accuracy of 98.21% and 0.0211 error rate. Fig. 3 shows the training and validation loss, and Fig. 4 shows the training validation accuracy.

Fig. 5 illustrates training accuracy and loss value of the model during train phrase. We have tested the model on test data, the model performance tested on the test set produced 97.37% accuracy as shown in Fig. 6.

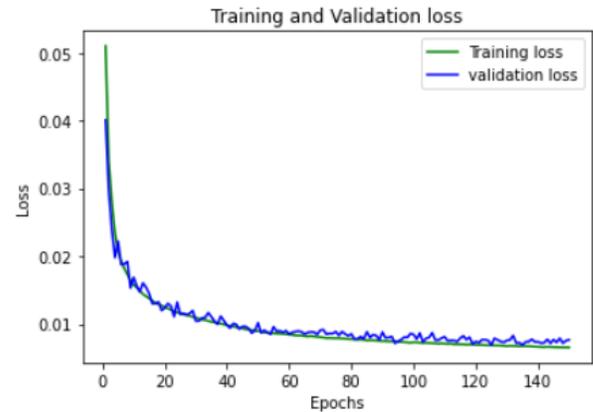


Fig. 3. Training and Validation Loss over 150 epochs.

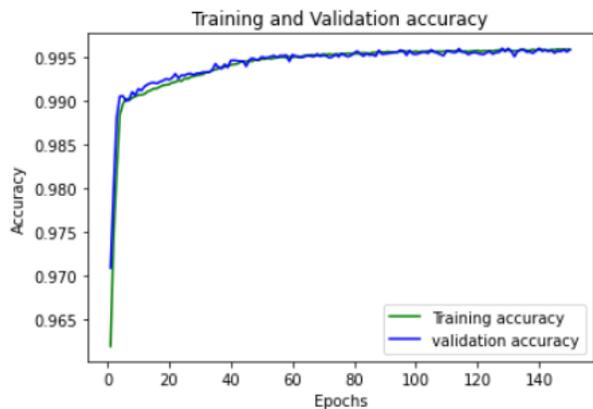


Fig. 4. Training and Validation Accuracy Graph over 150 epochs.

```

125973/125973 - 98s - loss: 0.0339 - accuracy: 0.9743
125973/125973 - 94s - loss: 0.0234 - accuracy: 0.9811
125973/125973 - 92s - loss: 0.0219 - accuracy: 0.9817
125973/125973 - 92s - loss: 0.0211 - accuracy: 0.9821
    
```

Fig. 5. Training Performance of our Model.

```

1656/1656 [=====]
value Loss: 0.030921630561351776
value Accuracy: 0.9736678004264832
    
```

Fig. 6. Test Performance of our Model.

V. RESULT

The NSL-KDD dataset, on the other hand, was utilized to test this approach. The proposed machine learning-based categorization solution for DDoS attacks has high accuracy in testing. From the results of the trained model, it was observed the proposed model's accuracy is 98.21%, which is almost a perfect method to prevent and protect the 22 different types of attacks, including DDoS attacks. Moreover, the LSTM evaluation model generates a 97.37% accuracy in the test set, the algorithm fits the patterns of the dataset with a 97.37% accuracy.

VI. COMPARISON OF RELATED WORK

Table III illustrates the comparison results of the accuracy between the proposed LSTM model with various deep learning methods using the same NSL-KDD dataset. Alkahtani et al. [22] conducted an experiment in which they chose the essential network features. To detect the anomaly in cybersecurity threats, these features were analyzed by classifying algorithms. SVM and KNN algorithms and deep learning based on the LSTM-RNN model were used to develop machine learning models. When compared to the KNN and LSTM, the SVM method produces better results. In the KDD Cup '99 and NSL-KDD datasets, the SVM method performed better than the LSTM-RNN and KNN methods. Their deep learning technique, based on the LSTM-RNN algorithm, had a high accuracy of 93.55%, but it couldn't surpass SVM's performance. Furthermore, they split the data into 70/30 train-test ratios in their experiment, while the proposed model utilized an 80/20 train-test ratio. As a result, the 80/20 split ratio produces better model accuracy than the 70/30 split ratio used in the LSTM-RNN algorithms.

TABLE III. ACCURACY COMPARISON FOR VARIOUS DEEP LEARNING TECHNIQUE WITH PROPOSED MODEL USING NSL-KDD DATASET

Authors	Technique	Accuracy testing model (%)
Alkahtani et al. [22]	LSTM-RNN. Support Vector Machine (SVM). K-Nearest Neighbor (K-NN).	93.55 (LSTM RNN) 96.53 (SVM) 87.65 (KNN)
Tang et al. [15]	Gated Recurrent Unit Recurrent Neural Network (GRU-RNN)	89.00
Niyaz et al. [23]	Self-taught Learning (STL), a deep learning-based technique	88.39
Ugwu et al. [11]	LSTM + SVD	90.59
Proposed model	LSTM-RNN	97.37

The "Adam" optimizer for DNN optimization was utilized in our study, which reduces the loss and optimizes the model. To transform order-like values to numeric numbers, the researcher uses the "OneHotEncoder" and "OrdinalEncoder" libraries. Despite their excellent performance, our proposed model has achieved better outcomes with a testing accuracy model of 97.37% and a loss value of 0.0287 from 52977 sample test packets. This evaluation reveals that the LSTM is

an effective solution for preventing and protecting against 22 different sorts of attacks.

The accuracy of our proposed method against the other approaches is significantly different in these comparisons. In the NSL-KDD dataset, our LSTM-RNN beats models that utilize all 41 features for training and testing. When compared to previously implemented deep learning methods in Alkahtani et al. [22], Tang et al. [15], Niyaz et al. [23], Ugwu et al. [11], the proposed model of LSTM-RNN did very well on the evaluation of the test data. This comparison demonstrates how our method's clear phases are predictable, accurate, effective, and authoritative.

Tang et al. [15] claim that when using a GRU-RNN method, their Deep Recurrent Neural Network (DNN) methodology obtained an accuracy of 88.39%. They use a Nadam optimizer and a mean squared error (MSE) model in their experiment. Our proposed model was developed using the Adam optimizer, which is the best optimizer. According to Kandel et al. [24], each optimizer is compared differently depending on the architecture, and the Adam optimizer has the best performance on the dataset in evaluation. They also used the mean square error (MSE) method to remove and appreciate the average error. Mean absolute error (MAE) was utilized in the proposed model. As a consequence, MSE outperformed the MSE technique in terms of interpretation.

Niyaz et al. [23] use NIDS based on sparse autoencoder and soft-max regression. As a result, they claim that the NSL-KDD dataset's Normal and anomaly (2-class) classification yielded an accuracy of 88.39%. By monitoring their method, autoencoder trains to effectively represent a manifold on which the training data resides. It was done by utilizing the mean square error (MSE) approach, which does not show an average error.

Ugwu et al. [11] designed the LSTM and SVD deep learning methods to show considerable improvement. They pre-processed the network data by converting features and normalizing the data. Non-numeric feature values were converted to numeric values using the feature conversion method. Their feature conversion method is nearly identical to the feature selection technique employed in our proposed model. However, our proposed model outperformed theirs.

VII. DISCUSSION

The machine learning detection approach was proposed and addressed in identifying a DDoS attack in this research study. This research has led to the understanding that many traditional machines learning, and deep learning methods can be used to detect a DDoS attack. However, when deciding whether to use traditional machine learning or deep learning with a large dataset, deep learning was considered due to its ability to solve difficult issues involving finding hidden patterns in data. It has a deep understanding of the complex relationships among a huge number of interdependent variables. Deep learning algorithms can create far more efficient decision rules. Deep learning is particularly effective in this study because the NSL-KDD dataset frequently requires dealing with unstructured data. Our findings reveal that LSTM is a nearly ideal strategy for preventing and protecting against 22 different types of

attacks. Classical machine learning, on the other hand, can be a preferable solution for smaller jobs that require less complex feature engineering and do not require the analysis of unstructured data.

VIII. CONCLUSION

The study has focused on presenting and demonstrating the design, implementation, and testing of a Detecting DDoS by Machine Learning solution to provide end-users with machine learning-based detection of DDoS attacks. End-users can re-route all traffic to an external server with DDoS mitigation capabilities hosted. The designed model solution allows researchers to build network-based detection models for network attacks using multiple machine learning methods, primarily classification. This result concludes that the objective to explore the type and study the characteristics of a DDoS attack from the viewpoint of machine learning was successfully achieved.

From the observation of the results for the proposed machine learning method, the LSTM RNN-based classification algorithm enhanced the detection of DDoS attacks. Pre-processing, attribute selection, and a detection and prevention system are the three components that the researcher proposes. The LSTM evaluation model fitting with the LSTM algorithm is demonstrated in the final phase. Significant testing was carried out, and the findings reveal that LSTM-RNN greatly surpasses existing DDoS attack detection systems. The algorithm learns the dataset's patterns with a 97.37% accuracy with a 0.0309 value loss. It was considerably easier to train numerous models in a short amount of time with TensorFlow, Google's second-generation machine learning framework. The LSTM recurrent neural network algorithm has been shown to have higher accuracy in detecting DDoS attacks in this study. These results covered achieving the objectives to improve DDoS attack detection approaches using a machine learning model to analyze network traffic activities for cyber threat detection; and to discover a feature selection strategy that, when combined with a machine learning system, can improve DDoS detection rates.

In a comparison of related work, the accuracy of our proposed method against all the other methods is significantly different. Our LSTM-RNN outperforms models that use all 41 features for training and testing in the NSL-KDD dataset. The proposed LSTM-RNN performed very well on the evaluation of the test data when compared to previously applied deep learning methods in Alkahtani et al. [22], Tang et al. [15], Niyaz et al. [23], Ugwu et al. [11]. This demonstrates how our method's phases are predictable, accurate, effective, and authoritative.

IX. FUTURE WORK

Even though DDoS packets or attacking packets are known, DDoS detection is not perfect. In the future, there will always be diverse approaches. However, this scope is not defined as one of the projects' objectives that should be achieved. A solution for improvement could be to add more datasets to the proposed system; another feature Selection technique that can be used; And other classifiers could be added to improve attack detection; Use of confusion matrix to show the mistaken type of attacks. In the future direction of this research, I suggest using advanced deep learning algorithms to build a predictive analytics model to

develop an automated system that can react based on current situations to analyze incoming data in networks. It could decide on defense mechanisms, evaluation and provide safety data on what is going on in a network.

REFERENCES

- [1] K. Singh, K. S. Dhindsa, and D. Nehra, "T-CAD: A threshold based collaborative DDoS attack detection in multiple autonomous systems," *J. Inf. Secure. Appl.*, vol. 51, p. 102457, 2020.
- [2] T. Ray, "DDoS defense: new tactics for a rising shadow industry," *Network. Secure.*, vol. 2020, no. 4, pp. 6-7, 2020.
- [3] S. Sambangi and L. Gondi, "A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression," *Proceedings*, vol. 63, no. 1, p. 51, 2020.
- [4] X. Yuan, C. Li, and X. Li, "DeepDefense: Identifying DDoS Attack via Deep Learning," *2017 IEEE Int. Conf. Smart Comput. SMARTCOMP 2017*, pp. 1-8, 2017.
- [5] A. Guerra-Manzanares, H. Bahsi, and S. Nomm, "Hybrid feature selection models for machine learning based botnet detection in IoT networks," *Proc. - 2019 Int. Conf. Cyberworlds, CW 2019*, pp. 324-327, 2019.
- [6] U. Sabeel, S. S. Heydari, H. Mohanka, Y. Bendhaou, K. Elgazzar, and K. El-Khatib, "Evaluation of Deep Learning in Detecting Unknown Network Attacks," *2019 Int. Conf. Smart Appl. Commun. Networking, SmartNets 2019*, 2019.
- [7] M. Rusyaidi and Z. Ibrahim, "A Review: An Evaluation of Current Artificial Intelligent Methods in Traffic Flow Prediction," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 917, no. 1, 2020.
- [8] M. S. Elsayed, N. A. Le-Khac, S. Dev, and A. D. Jurcut, "DDoSNet: A Deep-Learning Model for Detecting Network Attacks," *Proc. - 21st IEEE Int. Symp. a World Wireless, Mob. Multimed. Networks, WoWMoM 2020*, pp. 391-396, 2020.
- [9] F. O. Catak and A. F. Mustacoglu, "Distributed denial of service attack detection using autoencoder and deep neural networks," *J. Intell. Fuzzy Syst.*, vol. 37, no. 3, pp. 3969-3979, 2019.
- [10] J. D. Gadze, A. A. Bamfo-Asante, J. O. Agyemang, H. Nunoo-Mensah, and K. A.-B. Opare, "An Investigation into the Application of Deep Learning in the Detection and Mitigation of DDOS Attack on SDN Controllers," *Technologies*, vol. 9, no. 1, p. 14, 2021.
- [11] C. C. Ugwu, O. O. Obe, O. S. Popoola, and A. O. Adetunmbi, "A distributed denial of service attack detection system using long short term memory with Singular Value Decomposition," *Proc. 2020 IEEE 2nd Int. Conf. Cyberspace, CYBER Niger. 2020*, pp. 112-118, 2021.
- [12] Ö. KASIM, "An efficient and robust deep learning based network anomaly detection against distributed denial of service attacks," *Comput. Networks*, vol. 180, no. June, 2020.
- [13] Y. Gormez, Z. Aydin, R. Karademir, and V. C. Gungor, "A deep learning approach with Bayesian optimization and ensemble classifiers for detecting denial of service attacks," *Int. J. Commun. Syst.*, vol. 33, no. 11, pp. 1-16, 2020.
- [14] M. D. Hossain, H. Ochiai, D. Fall, and Y. Kadobayashi, "LSTM-based Network Attack Detection: Performance Comparison by Hyperparameter Values Tuning," *Proc. - 2020 7th IEEE Int. Conf. Cyber Secur. Cloud Comput*, pp. 62-69, 2020.
- [15] T. A. Tang, D. McLernon, L. Mhamdi, S. A. R. Zaidi, and M. Ghogho, "Intrusion detection in sdn-based networks: Deep recurrent neural network approach," *Adv. Sci. Technol. Secur. Appl.*, pp. 175-195, 2019.
- [16] A. O. Alzahrani and M. J. F. Alenazi, "Designing a network intrusion detection system based on machine learning for software defined networks," *Futur. Internet*, vol. 13, no. 5, 2021.
- [17] P. Nimbalkar and D. Kshirsagar, "Feature selection for intrusion detection system in Internet-of-Things (IoT)," *ICT Express*, vol. 7, no. 2, pp. 177-181, 2021.
- [18] W. Nazih, Y. Hifny, W. S. Elkilani, H. Dhahri, and T. Abdelkader, "Countering ddos attacks in sip based voip networks using recurrent neural networks," *Sensors (Switzerland)*, vol. 20, no. 20, pp. 1-15, 2020.
- [19] R. C. Staudemeyer and E. R. Morris, "Understanding LSTM -- a tutorial into Long Short-Term Memory Recurrent Neural Networks," pp. 1-42, 2019.

- [20] F. E. Laghrissi, S. Douzi, K. Douzi, and B. Hssina, "Intrusion detection systems using long short-term memory (LSTM)," *J. Big Data*, vol. 8, no. 1, 2021.
- [21] S. Althubiti, W. Nick, J. Mason, X. Yuan, and A. Esterline, "Applying Long Short-Term Memory Recurrent Neural Network for Intrusion Detection," *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2018-April, 2018.
- [22] H. Alkahtani, T. H. H. Aldhyani, M. Al-Yaari, and M. Y. Alzahrani, "Adaptive Anomaly Detection Framework Model Objects in Cyberspace," 2020.
- [23] Q. Niyaz, W. Sun, A. Y. Javaid, and M. Alam, "A deep learning approach for network intrusion detection system," *EAI Int. Conf. Bio-inspired Inf. Commun. Technol.*, 2015.

Proficient Networking Protocol for BPLC Network Built on Adaptive Multicast, PNP-BPLC

Ali Md Liton, Zhi Ren, Dong Ren, Xin Su

School of Information and Communication Engineering Chongqing
University of Posts and Telecommunications, Chongqing, China

Abstract—In order to solve the problems of the existing broadband power line carrier communication standard IEEE1901.1 data link layer protocol network, multiple primary nodes receiving the beacon will send connotation entreaty message, and CCO will send an association sanction message proximately after receiving the message to confirm their character the central coordinator CCO association confirmation message reply is not timely to reduce the success rate of network access, high network access delay and control overhead. Based on the characteristics of BPLC carrier network, Proficient networking instrument of broadband PLC built on adaptive multicast PNP-BPLC is proposed in this paper. The simulation results show that adaptive multicast is used when CCO replies to the primary station, which can excellently recover the success rate of low-voltage PLB carrier communication nodes, decrease the network access delay and cut the network control overhead. Finally we used to OPNET simulation software to prove simulation result.

Keywords—Powerline communication; network delay; control overhead; central coordinator

I. INTRODUCTION

Low voltage broadband powerline carrier communication (BPLC) [1, 2, 3] is a communication mode that uses low-voltage power distribution line (380/220V subscriber line) as information transmission medium for voice or data transmission. It has the natural advantage that there is no need to reset up the network. The BPL carrier communication has wide bandwidth, and the basic frequency band is 1MHz ~ 20MHz Compared with the traditional NBPLC carrier communication, Broadband PLC has higher transmission rate, stronger anti-interference performance and better performance. At present, the international standard for broadband power line carrier communication is IEEE 1901.1 [4-5]. This standard is created on Q/GDW 11612 technical description for interconnection and interworking of low voltage PLB carrier communication [5] of the state grid corporation of china, and grasps the operative tender of internet of things expertise based on power line carrier communication in energy internet [6]. BPLC has been widely used in automatic meter reading [7 8 9], intelligent power consumption system [8], street lamp control [8], charging pile construction [10 11], etc. At present, there are a lot of research in the field of BPLC technology. For example, literature [10, 11, 12] explores the features of low-voltage BPL carrier communication channel. For the input impedance features, signal noise and interloping characteristics, phase shift characteristics, the actual communication area of PLC is hundreds of meters. In order to

confirm link stability and long distance transmission, literature [13-14] offers that tree networking and multi hop transmission similar to wireless sensor networks are assumed in the network management sublayer. Many people have planned the BPLC networking algorithm Literature [14-15] proposed an improved Q-learning algorithm suitable for multi restraints of PLC, LAN through incessant collaboration with the unknown situation. Although this kind of ant colony algorithm has certain adaptive ability, it needs to send a large number of control messages to interact with neighbor nodes, which has the problems of excessive control overhead and waste of resources. Reference [16-17] provides a tree based networking method for low voltage BPL carrier communication network, which uses topology assessment frame to travel the network topology, upholds and updates the sub node set according to the lowest energy value required for normal [18,19,20] communication, picks candidate routing nodes in turn according to the energy value, and then assigns [24-25] network address to each sub node in the sub node set, Although this technique is simple and easy, it does not give the vigorous rebuilding and optimization manner of routing, and the network constancy is poor. IEEE1901.1 [21, 22, 23,] standard also espouses impart networking. The CCO elicits the network access appeal of STA level by level by sending the central beacon, placing the sending of discovery beacon and sending of proxy beacon to ample the entire networking procedure.

Rest of the paper we have described simulation results of the paper. In the Section of III(B)(1) we described, Network Access success rate, in Section III(B)(2) Average network access delay and in Section III(B)(3) Network control overhead, proposed PNP-BPLC Protocol is better than the other two protocols.

II. SYSTEM MODEL AND PROBLEM DESCRIPTIONS

A. Network Scenario

The broadband carrier communication network will normally form a multi-layer tree network with the central coordinator CCO as the center and the proxy coordinator PCO as the relay agent to connect all stations STA. As shown in Fig. 1, the topology of a typical BPL carrier communication network is shown.

B. Protocol Stack Structure of BPLC

Based on the standard open system interconnection (OSI) seven layer model, the Broadband carrier communication network protocol stack defines three layers, physical layer, data link layer and application layer.

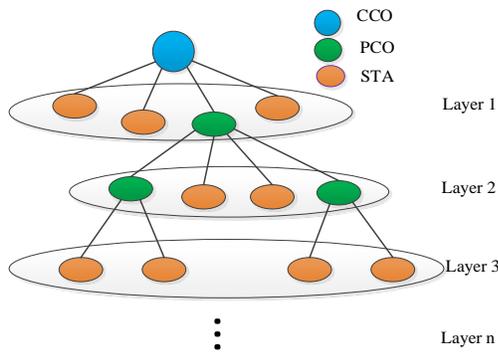


Fig. 1. Topology of BPL Carrier Communication Network.

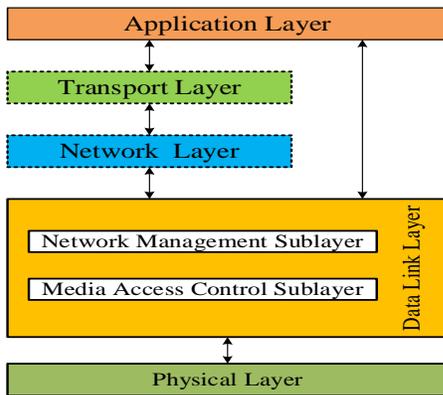


Fig. 2. BPLC Communication Network Protocol Stack.

The basic structure is shown in Fig. 2. The data link layer is divided into network management sublayer and media access control sublayer. The data link layer directly provides transmission services for the application layer and can also be prolonged to edge with standard TCP / IP to grasp standard IP network communication.

C. The BPLC Networking

In order to certify the consistency of BPLC network routing and data spread, BPLC network will be networked. The procedure is as follows.

- 1) After CCO is powered on, coordinate the time slot and network documentation among networks. After effective organization, demeanor solitary network networking.
- 2) CCO starts to send the central beacon in the beacon slot, and the station receiving the central beacon wants to enter by sending the association appeal message in the CSMA slot access.
- 3) CCO verifies the site requesting network access through the white list. After positive confirmation, CCO sends the dispensation result to the STA site requesting network entree through the connotation sanction message, expressive the effective network access of the node.
- 4) After the primary node effectively arrives the network, CCO will arrange beacon time slot for it to send discovery beacon. The sending of discovery beacon can elicit the secondary site around the new site to recruit the request for related network access.

- 5) This rotation allows the main level STA site utmost from CCO to connection the network.

D. Problem Description

It is found that there are two problems in the BPLC networking progression.

1) In BPLC networking, after CCO sends the central beacon in the beacon slot, multiple primary nodes receiving the beacon will send connotation entreaty message, and CCO will send an association sanction message proximately after receiving the message to confirm their character. With the STA node layer by layer forwarding the association request message network entire tender through PCO, the number of STA nodes increases, and the association authorization message is likely to be hugged in the CCO queue, ensuing in the CCO inept to apportion TEI, time slot and other evidence to the nodes to be edited in time, and the network access solicitation node cannot access the network normally, resulting in the reduction of the network access attainment rate of BPLC network nodes.

2) STA sends a huge number of connotation entreaty messages, and CCO needs to send Association validation messages of the similar scale for retort. However, distinct beacon time slots, CSMA time slots assign superior time slots for separately node, so encounters must be dodged, ensuing in CCO's fiasco to reply relationship validation messages to overtone request messages sent by some nodes smearing for network access in time, then these STA,s will prompt the instrument of resending connotation request message because they cannot collect overtone sanction message within the waiting time edge T_{max} , consequent in the rise of control overhead and network access delay.

E. PNP-BPLC Mechanism

In order to solve the problems described in the previous section, this paper offers a BPLC proficient Networking (PNP) mechanism built on adaptive multicast, including two new mechanisms adaptive multicast reply Association confirmation message and association reply based on retransmission message. The specific ideas are as follows.

F. Adaptive Multicast Reply Connotation Confirmation Message Mechanism

For the problem a) declared in the previous section, seeing that the BPLC is a tree topology, some STA sub sites will be equestrian under PCO, which means that the higher the level is, the more stations are likely to be. Then it is likely that the association confirmation messages sent by multiple websites to be accessed will be put into the cache queue by CCO and not replied in time. Therefore, this paper propositions CCO adaptive multicast replies Association sanction message apparatus.

The core idea is CCO queries the number of association confirmation messages in its queue. When the number is 1, unicast Association confirmation messages. When the number is greater than 1, multicast association instant warning messages are sent to send association reply messages more

rationality, which increases the network access success rate of nodes and condenses the network access delay and control overhead. Since multilevel nodes will also forward the association request message to the prime node, which is principally the same as that of the prime node, this paper mainly discusses the association retort of CCO to the principal node. The basic process is as follows.

Step 1: Start the neighbor network observing timer when CCO is drove on. If the inter network direction frame is received within the listening time t , coordinate the inter network identification NID and time slot. If the inter network direction frame is not received within the listening time t , single network networking is carried out. CCO recordings the central beacon in the beacon slot.

Step 2: The neighbor node receives the central beacon, forms whether the "start association flag bit" of the beacon is 1, and is ready to send the association entreaty message, before the station is ready to send, listen for data being transmitted on the bus whether the line is busy or not. If the line is found to be busy during listening, wait for a delay and listen again. If it is still busy, continue to delay the wait. If the waiting times n exceed the threshold 16, STA retransmits the association entreaty message and marks the message. If the time of each delay is erratic, it is resolute by the reduced two the M-ray exponential bakeoff algorithm fixes the delay value.

Step 3: CCO receives the association request message sent by the node relating for network access, inquiries the STA site information consistent to the association request message, and completes white list confirmation. CCO queries the number of layers of the site and regulates the terminal address of the next hop. If it is a first class site, the destination address is the node. If it is a multi-layer site, the destination address is the PCO of the lowest level agreeing to the inviting site. CCO allocates TEI and time slot to the node and keeps them in the reminder reply message. CCO makes association reply to the principal site. If the channel is busy when sending Association reply, CCO will put the association reply message into a distinct association reply backlog.

Step 4: Check the number of association retorts in the queue. If it is 1, wait for the channel idle unicast to send the association sanction message. If it is greater than 1, all overtone messages in the queue are taken out, the address evidence of each node is found, and the association rapid hint message is formed. When the channel is idle, multicast is sent to the crucial node.

Step 5: The primary node limits whether it is the node applying for network access. If so, the network entree is effective. If not, the node is the lowest level PCO of the node applying for network access. The node forms an association endorsement message through the association reply message sent by CCO to itself and sends it to the next level node. The next level node also benches whether it is the node applying for network access first. If not, the node is the lowest level PCO of the node applying for network entree.

Then the node is the secondary PCO of the network access node. The node installs the address evidence and remains to send the association sanction message to its subsidiary nodes,

and so on until the association request node is found and the network entrée is effective.

G. Association Reply Apparatus Based on Retransmission Message

In order to solve the problem b), this paper propositions an association reply apparatus based on retransmission message. Its core idea is perceptive.

CCO adaptively sends the first level reply message by arbitrating whether the retransmitted association request message has been received around. If CCO receives the association request message retransmitted from STA node, it specifies that the association request message formerly sent by STA node has not received Association confirmation, which tortuously indicates that the BPLC network channel is busy.

If CCO remains to unicast Association credit messages in CSMA time slots, the channel cramming will be impaired. CCO stops generating association confirmation message unicast transmission and selects to verify the received association request message, and generates association rapid hint message, so that there is no large number of association confirmation messages competing with association request messages in CSMA time slot, so as to slow down the message load of the channel and reduce the control overhead.

Fig. 4 is 1 graphic diagram of association reply apparatus built on retransmission message. Fig. 4(1) shows the uplink association request message sent by STA and the association sanction message sent by CCO under normal conditions. Fig. 4(2) shows the graphic diagram of more network access nodes. As can be seen from the figure, due to the large number of nodes, both uplink and downlink messages need to compete for channels, resulting in the STA sending the uplink Association reply message represented by the blue arrow not receiving the agreeing association confirmation message and not being able to access the network normally. After a waiting time, the association request message with the green arrow is sent again. Fig. 3 is the graphic diagram of the new apparatus, and the red arrow represents the association instant indication message sent by CCO multicast. In Fig. 3, after CCO senses that some STA resend association request messages, it uses the method of ending sending Overtone sanction messages and multicast Overtone instant indication messages, which almost reduces the control overhead by 50% and importantly cuts the drain of the channel.

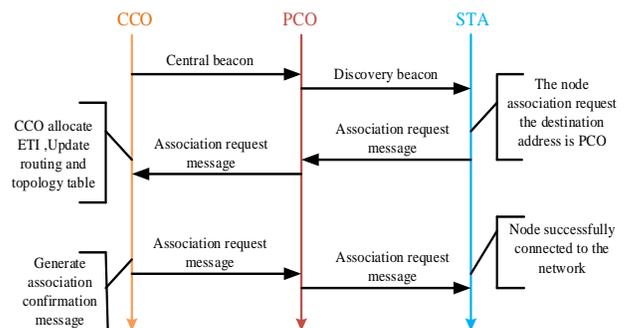


Fig. 3. Message Interaction of Associated Network for BPLC.

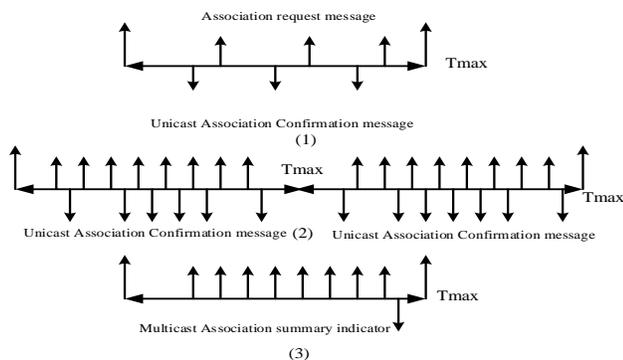


Fig. 4. Association Reply Apparatus based on Retransmission Message.

The Basic Operation Process is as follows:

Step 1: CCO performs network organization, which is reliable with CCO in the adaptive multicast reply association confirmation message device.

Step 2: STA makes association request, which is consistent with the association request made by STA in the adaptive multicast reply Connotation confirmation message apparatus.

Step 3: CCO receives the association request message sent by STA and benches whether there is a retransmission flag. If so, it stops sending the connotation confirmation message. In this time slot, it only accepts the association request message, takes out the node ID, original address and other information, and forms an association instant hint message. Multicast sends Association instant indication message at T_{max} time. The algorithm of multicast sending association rapid hint message is consistent with that of adaptive multicast reply connotation confirmation message apparatus.

Step 4: The main node regulates whether it is the node smearing for network access. If so, the network access is effective. If not, the node is the lowest level PCO of the node smearing for network access, forming an association confirmation message. This procedure is constant with the progression in which the node judges whether it is the node smearing for network access in the adaptive multicast reply association confirmation message apparatus.

H. PNP-BPLC Apparatus Progression

The flow chart of BPLC proficient networking apparatus based on adaptive multicast is shown in Fig. 5, which includes two new mechanisms adaptive multicast reply association confirmation message and association reply built on retransmission message.

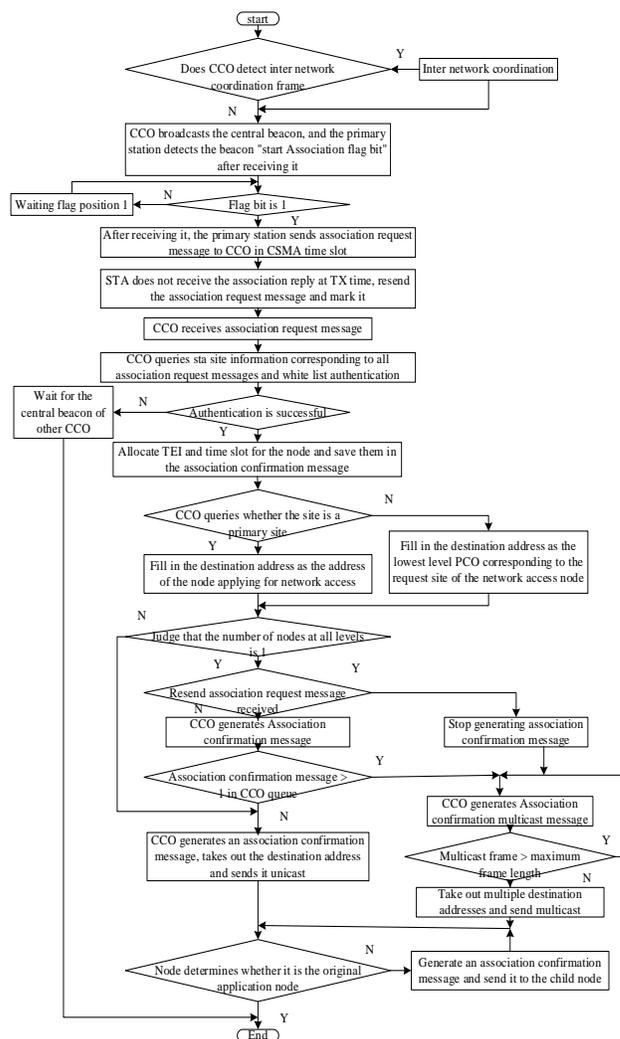


Fig. 5. Adaptive Multicast Access Protocol Process.

III. SIMULATION ANALYSIS

A. Simulation Parameter Setting

This paper uses OPNET 14.5 simulation tools to simulate IEEE1901.1 technical standard and Literature [15], PNP-BPLC apparatus are used to simulate the network access success rate, average network access delay and control overhead. The main simulation parameters are shown in Table I.

TABLE I. MAIN PARAMETER SETTING

Simulation Parameter	Value
Simulation scene /m*m	3000*1000
Simulation running times/s	100
Number of nodes/Pc	[4,9,14,19,24]
Transmission rate /Mbps	10
Single level network access time/s	20
Single hope effective transmission distance /M	500
Number of CCO	1
Maximum hops/hop	3

B. Performance Index

1) *Network access success rate*: Indicates the network access success rate of each STA site of BPL carrier communication network. The network access success rate P is calculated as Formula 1.

$$P = \sum_{i=1}^{i=n} Xi \div \sum_{i=1}^{i=n} Yi \times 100\% \tag{1}$$

Where n represents the simulation scenario with n nodes, Xi represents the number of association request messages sent by the *i*th node, and Yi represents the number of association confirmation messages received by the *i*th node.

2) *Average network access delay*: The average network access delay represents the generation time T_{pk} from the creation of the central beacon frame the average time of T_{sim} when the association reply message sent by the create to the CCO is received by the node applying for network access. With transmission delay, spread delay, inter layer processing delay and MAC layer queuing delay. The average network access delay is calculated according to formula 2.

$$\bar{T} = \frac{1}{n} \sum_{i=1}^{i=n} (T_{sim}(i) - T_{pk_creat}(i)) \tag{2}$$

3) *Network control overhead*: Network Control overhead refers to the sum of bits of control message required in the whole networking process of BPLC network. In BPLC network, there are four types of control messages inter network coordination message, beacon message, association request message and association reply message. Because the simulation is in a single network environment, the overhead message in this paper does not include inter network coordination message. Among them, the beacon message is divided into central beacon, proxy beacon and discovery beacon. The association reply message also includes association validation message and association summary indication message. The calculation of BPLC control cost C is shown in Formula 3.

$$C = 8 * (M * (c_{11} + c_{12} + c_{13}) + N * c_2 + O * (c_{31}) + P(c_{32})) \tag{3}$$

Among them, C11, C12, C13, C2, C31 and C32 are the number of central beacon, proxy beacon, discovery beacon, overtone request message, association confirmation message and association rapid indication message singly, which can be counted by simulation. M, N, O and P, are beacon frames, association request message, association validation message and association instant indication message, respectively. According to IEEE 1901.1 standard, MAC frame is the basic transmission unit for transmission between MAC layers of different stations. A MAC frame consists of MAC frame header, MAC service data unit MSDU and integrity check value. In this paper, the size of long frame header is h = 26 bytes and the size of integrity check value is w = 4 bytes.

The beacon frame size M is calculated according to formula 4.

$$M = h + m + w \tag{4}$$

M is beacon frame MSDU size = MPDU frame control size 16 bytes + frame load 520 bytes = 536 bytes. Thus, M = 576 bytes can be obtained.

The calculation of association request message n is shown in formula 5.

$$N = h + n + w \tag{5}$$

N is the MSDU size of association request message = 4 bytes of Association message header + 64 bytes of association request message format = 68 bytes. It can be concluded that n = 98 bytes.

The calculation of association confirmation message O is shown in formula 6.

$$O = h + o + w \tag{6}$$

O is the MSDU size of association confirmation message = 4 bytes of association message header + 64 bytes of association confirmation message format = 68 bytes. It can be concluded that o = 98 bytes.

The calculation of association request message P is shown in formula 7.

$$P = h + p + w \tag{7}$$

P is the size of MSDU of association instant hint message = 4 bytes of header of association message + format of association summary indication message q, q = basic length of association instant indication message + extended variable length.

It is assumed that the maximum threshold of association confirmation messages queued in CCO queue at the same time is 10. Therefore, q is the maximum number of fills 10 * each site information field, as shown in Table II, the size is 8 bytes = 80 bytes. Therefore, P = 84 + 30 = 114 bytes.

TABLE II. SITE INFORMATION FIELD

Field	Byte number	Bits	Field size	Definition
Site address 1	0-5	0-7	6 bytes	MAC address
Site ETI 1	6	0-7	12 bits	TEI assigned to site
Site ETI 1	7	0-3	Site TEI 1	TEI assigned to site
Retain	7	4-7	4bits	Retain
.....
Site n address	Site n address
Site TEI n	TEI n=Total number of stations

It can be proved theoretically that the message overhead using PNP apparatus is less than that of the original BPLC network access apparatus.

Condition (1): Except for the associated reply message, the overhead of other control messages is equal and is not counted

Condition (2): Number of messages in CCO Association confirmation queue $m > 1$.

Prove: The message overhead of PNP apparatus is less than that of the original BPLC network access apparatus.

Multicast association summary indication message overhead $C2 = 8 * (H + 8 * m + W) = (84m + 240)$ bit using the PNP mechanism; because $m > 1$; Therefore, $C1 - C2 = 784m - (84m + 240) = 700m - 240 > = 460\text{bit} > 0$; Get a certificate.

Using the original PLC networking apparatus, unicast association confirmation message overhead $C1 = 8 * (m * (H + O + W)) = 784m\text{bit}$;

C. Analysis of Simulation Results

1) *Network access success rate*: As shown in Fig. 6, the PNP-BPLC contrivance and IEEE1901.1, Literature [15] simulation comparison diagram of data link layer protocol data network access success rate. It can be seen that with the increase of the number of stations and network level, the network access success rate of the two mechanisms gradually decreases, which is due to the increase of channel load and the repeated retransmission of association request message through the comparison of the two figures.

It can be seen that when there are few nodes, the network access success rates of the two mechanisms are basically the same. With the increase of the number of nodes, the network access success rate of PNP device is higher than that of IEEE190.11 and Literature [15] network access appliance of 1 MAC protocol. Because the BPLC network is a multi-level and multisite network, the two improved methods proposed by PNP-BPLC contrivance successfully improve the network access success rate of BPLC network.

2) *Average network access delay*: As shown in Fig. 7, the PNP-BPLC device and IEEE1901.1, Literature [15] comparison diagram of network access delay of different node scenarios in 100s simulation time of 1 MAC layer protocol network access mechanism. Through comparison, it can be seen that with the increase of the number of nodes, the amount of data in the network gradually increases, resulting in the gradual increase of the average network access delay of nodes. The network access delay of PNP-BPLC appliance is lower than IEEE1901.1, Literature [15] the network access delay of 1 MAC layer protocol network access mechanism, and the advantages become more obvious with the number of nodes. Description in IEEE1901.1, Literature [15] under the standard, with the increase of the number of nodes, many association confirmation messages are not sent in time in the CCO queue. It increases the network access delay. The PNP-BPLC

contrivance reduces the transmission of control messages, reduces the congestion of the channel almost.

3) *Network control overhead*: As shown in Fig. 8, control overhead comparison diagram shows the PNP-BPLC and the original PLC network access mechanism IEEE1901.1, Literature [15] the comparison diagram of the control overhead of the standard network access mechanism can clearly see that the control overhead of the two apparatuses increases steadily when the BPLC enters the network step by step, because with the increase of nodes, more nodes need to send control messages, which upsurges the control overhead. Through comparison, it is obvious that the control overhead of PNP-BPLC appliance is basically the same as that of the original PLC network access mechanism when the number of nodes is small. This shows that when the number of nodes is small and the channel is idle, the reply of CCO to the network access node is mainly association confirmation message. However, in the multi-layer PLC network environment with a large number of nodes, the control overhead of the original PLC network access mechanism increases almost exponentially, while the control overhead of PNP-BPLC mechanism increases only linearly.

Depiction in IEEE 901.1, Literature [15] standard, with the growth of the number of nodes, many association confirmation messages are not sent in time in the CCO queue, resulting in retransmitted association request messages, which greatly increases the control overhead of the network. The PNP-BPLC mechanism can adaptively send the association reply message, which greatly reduces the overhead of the association confirmation message sent by the original CCO. At the same time, it also reduces the retransmission of the association request message by the network access application node, and reduces a large number of message overhead.

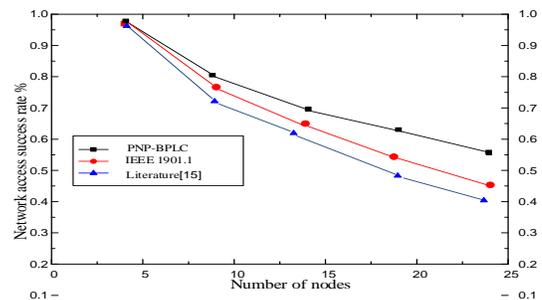


Fig. 6. Comparison of Success Rate of Network Access.

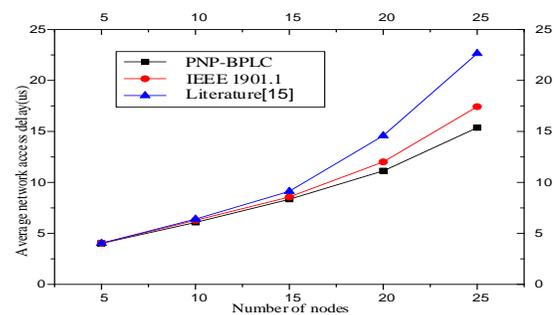


Fig. 7. Comparison of Average Network Access Delay.

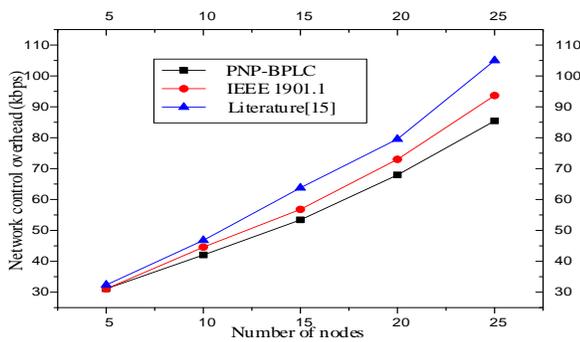


Fig. 8. Comparison of Network Control Overhead.

IV. CONCLUSION

The low voltage PLC communication network is affected by robust interference, which requires the networking process to have high adaptive ability. Aiming at the two problems of BPLC data link layer networking, combined with the BPLC network scenario, this paper propositions proficient BPLC networking apparatus built on adaptive multicast. It includes two new mechanisms adaptive multicast reply association confirmation message and association reply based on retransmission message. By sending the primary association reply message more reasonably, the channel resources are more reasonably utilized, and the BPLC networking efficiency is higher. The experimental results show that the above mechanism can improve the data access success rate of BPLC network, reduce the access delay, and greatly reduce the control overhead of multi-layer BPLC network, which verifies the effectiveness of PNP-BPLC mechanism.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61971080), the Key Project of Fundamental and Frontier Research of Chongqing (cstc2015jcyjBX0085), and the cooperative project of Chongqing Electricity Research Institute of State Grid of China (SGCQDK00NYJS1900346).

REFERENCES

- [1] Zhang Hao. Research on data link layer protocol simulation and performance of broadband power line communication system [D]. Xi'an University of Electronic Science and technology, 2017.
- [2] H. Hrasnica, A. Haidine, R. Lehnert, "Broadband Powerline Communications Network Design," Wiley, c2004. 275 s. ISBN 0-470-85741-2.
- [3] Galli S, Scaglione A, Zhifang W. For the grid and through the grid: the role of power line communications in the smart grid. P IEEE 2011; 99: 998-1027.
- [4] Massaki K. Introduction to robust, reliable and high-speed power-line communication systems. IEICE T Fund Electr 2001; 12: 2958-2965.
- [5] High Definition Power Line Communication (HD-PLC), [Online]. Available: <http://www.hd-plc.org/>
- [6] UPA Digital Home Specification (DHS), [Online]. Available: www.upapl.org

- [7] Guzelgoz, S. (2011). Characterizing wireless and power line communication channels with applications to smart grid networks. PhD Dissertation, University of South Florida.
- [8] IEEE Standard 1901™. (2010). IEEE Standard for Broadband over Power Line Networks: Medium Access Control and Physical Layer Specifications.
- [9] L. Zhang, X. Liu, and D. Xu, "A novel security monitoring system of coal mine based on power line communication dynamic routing technology," in Industry Applications Society Annual Meeting, 2014
- [10] S. Galli and T. Lys, "Next generation narrowband (under 500 kHz) power line communications (PLC) standards," China Commun., vol. 12, no. 3, pp. 1-8, 2015.
- [11] Chen Feng, Zheng-Wengang, Shen Chang-jun, et al. Low voltage power line carrier communication technology and application [J]. Power system protection and control, 2009,37 (22): 188-195.
- [12] IEEE Standard for Medium Frequency (less than 12 MHz) Power Line Communications for Smart Grid Applications," in IEEE Std 1901.1-2018, vol., no., pp.1-192,14May2018doi:10.1109/IEEESTD.2018.8360785.
- [13] Q / GDW 11612-2016, technical specification for interconnection and interworking of low voltage power line broadband carrier communication [S].
- [14] Fan yun-nian. Energy Internet, Qugao is not harmonious [n]. China Science Daily, April 23, 2020 (006).
- [15] Duke, Wang Wei-jie, Mei Ping. Application of OFDM based low-voltage power line broadband carrier communication technology in low-voltage centralized reading system [J]. Technology outlook, 2016,26 (35): 83-84
- [16] Lv Wei-jia. Application analysis of broadband power carrier technology in intelligent power system [a]. Proceedings of 2017 smart grid development seminar [C]. China Electric Power Research Institute: Taiji Computer Training Center, Haidian District, Beijing, 2017:5.
- [17] Tong Shi-wei. Research on channel characteristics test and modeling of broadband power line based on street light line [D]. North China Electric Power University, 2019.
- [18] Tan Bao-qi. Application of PLC technology of power carrier communication in the construction of charging pile [J]. China Equipment Engineering, 2019 (22): 161-162.
- [19] [28] Yan Yuan-zhi. Analysis of channel characteristics of low voltage power line carrier communication [J]. Communication world, 2018 (08): 160-161
- [20] Pan Dong-yang. Research on OFDM based low-voltage power line carrier communication system [D]. Chongqing University of technology, 2019.
- [21] Zhang Jie. Research on Routing Technology Based on broadband power line carrier communication [a]. Proceedings of 2018 smart grid information construction seminar [C] of China Electric Power Research Institute. China Electric Power Research Institute: the Sixth Research Institute of China Electronic Information Industry Group Co., Ltd., 2018:4.
- [22] Guo pan.Methods,devices,equipment, concentrators and systems for obtaining routes [P]. Cn102970233a, March 13, 2013.
- [23] Li Gui-lin, Wei Sheng-qing. A tree based networking method for low-voltage power line broadband carrier communication [P]. Cn107332777a, 2017-11-07.
- [24] Long Yu-li. Form of relay network for power line carrier communication in low voltage distribution network [J]. Electronic technology and software engineering, 2018 (20): 41.
- [25] Guangyu Pei and C. Chien. Low Power TDMA in Large Wireless Sensor Networks. In Military Communications Conference, 2001. MILCOM 2001. Communications for Network-Centric Operations: Creating the Information

Method for Improvement of Ocean Wind Speed Estimation Accuracy by Taking into Account the Relation between Wind Speed and Wind Direction

Kohei Arai¹, Kenta Azuma²

Science and Engineering Faculty, Saga University, Saga City, Japan¹
Former Student, Saga University, Saga City, Japan²

Abstract—A method for improvement of ocean wind speed estimation accuracy by taking into account the relation between wind speed and wind direction is proposed. Brightness temperature observed with microwave radiometer onboard satellite is modified with microwave radiometer derived wind direction proposed by Frank Wentz. Using the modified brightness temperature, more precise wind speed is estimated. Experiments with AMSR-E and NCEP GDAS data show improvements of wind speed estimation in comparison to the existing method based on the geophysical model of Frank Wentz together with the retrieval algorithm of Akira Shibata.

Keywords—Sea surface wind speed (WS); advanced microwave scanning radiometer for earth observing system (AMSR-E); NCEP Global Data Assimilation System (GDAS); relative wind direction (RWD)

I. INTRODUCTION

The physical quantity can be estimated. This is called remote sensing. Remote sensing generally has visible, near-infrared, thermal-infrared, and microwave observation wavelengths, and sensors are used [1]. Compared to other wavelengths, remote sensing that measures microwaves has poor resolution and can only make rough measurements [2]. However, it has the advantage of being observable even if it is affected by clouds, regardless of day or night. Furthermore, at present, the wind speed on the ocean surface can be remotely sensed only by sensors in the microwave wavelength range [3].

However, it is known that the estimation accuracy of the wind speed changes due to the influence of the relative wind direction in the marine physical quantity cage determination using microwaves. Frank. Wentz (2002) [4] describes the microwave radiation transfer equation and the marine physical quantity using the microwave radiation transfer equation when using the Top of Atmosphere (TOA) obtained from the microwave radiometer AMSR. We created an equation to estimate. Observation of atmosphere, ocean, and land with multi-wavelength microwave radiometer is introduced [5]. At this time, the influence of the relative wind direction was also taken into consideration. However, this effect was used in the microwave radiometer SSM / I and cannot be expected in AMSR. This is because AMSR and ASMR / I have different observation frequency bands.

Arai and Sakakibara proposed a method using the Wentz algorithm using improved simulated annealing for the radiometer ASMR-E [6]. This made it possible to simultaneously estimate various marine physical quantities, which led to an improvement in the estimation structure. On the other hand, Konda, Shibata, Ebuchi, and Arai (2006) used the estimated wind speed obtained from the AMSR product mounted on the observation satellite ADEOS-II and the wind direction obtained from the microwave scatterometer, SeaWinds onboard on the satellite [7].

One of the important issues of the improvement of ocean wind speed estimation accuracy is how to take into account the influence due to wind direction dependency from the wind speed estimation. This is the issue of this paper. The influence of the relative wind direction on the wind speed estimation was investigated, and the index of the influence of the relative wind direction under various situations was derived. They confirmed that the estimation accuracy was improved by performing the estimation using the derived index in the wind speed cage determination of ASMR and ASMR-E. However, in the case of simultaneous estimation of physical quantities, a method of modifying only the wind speed after estimation causes a contradiction in terms of estimation means. Therefore, it is desirable to influence the relative wind direction on the brightness temperature for simultaneous estimation. Therefore, in this study, the effect of relative wind direction is used using the global meteorological data (GDAS) obtained from the US National Environmental Forecast Center NCEP and the observed brightness temperature of the microwave radiometer AMSR-E mounted on the meteorological satellite aqua [8]. By analyzing the above, we analyzed how the relative wind direction affects the brightness temperature. Furthermore, by approximating the influence of the relative wind direction with several equations, we created a model that can derive the influence due to wind direction by linear calculation. By modifying Wentz's algorithm for estimating marine physical quantities with the model, the accuracy of marine wind speed estimation was improved. Experiments with AMSR-E and NCEP GDAS data show improvements of wind speed estimation in comparison to the existing method based on the geophysical model of Frank Wentz together with the retrieval algorithm of Akira Shibata.

The following section describes related research works followed by theoretical background. Some experiments are described. After that conclusion is described together with some discussions.

II. RELATED RESEARCH WORK

Simplified expression of the radiative transfer equation in thermal infrared window spectrum is proposed [9]. Evaluation of vector winds observed by NSCAT in the seas around Japan is conducted [10]. Polarization sensitivity of the ocean surface together with wind vector derived from POLDER and NSCAT on ADEOS is also evaluated [11].

Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing is conducted and validated with the truth data [12] together with simultaneous estimation of sea surface temperature, wind speed and water vapor with AMSR-E data based on improved simulated annealing [13].

Space and time retrieval of tide wind speed and wave height with altimeters onboard satellites based on PostGIS system is conducted [14]. Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing is also conducted [15].

Correction of the effect of relative wind direction on wind speed derived by AMSR is conducted [16]. Data fusion between microwave and thermal infrared radiometer data and its application to skin sea surface temperature, wind speed and salinity retrievals are proposed [17].

Comparative study of optimization methods for estimation of Sea Surface Temperature: SST and Ocean Wind: OW with microwave radiometer data is conducted [18].

III. THEORETICAL BACKGROUND

Effect of relative wind direction is considered based on the physical background. It is known that the brightness temperature of the upper atmosphere of microwaves largely depends on the gradient distribution of the sea surface and the state of the sea surface.

When the wind blows on the sea surface, the gradient distribution of the sea surface changes, and the brightness temperature at the upper end of the atmosphere fluctuates greatly. By utilizing this action, the sea wind speed can be estimated from the observed brightness temperature of a microwave radiometer or a microwave scatterometer. It is also known that the gradient distribution is distorted in the wind direction due to the wind.

It is considered that the upper-atmospheric brightness temperature changes further due to the distortion of the gradient distribution, that is, the upper-atmospheric brightness temperature also depends on the difference between the wind direction and the observation direction. The difference between the azimuth of the wind and the azimuth of the observation is called "Relative Wind Direction (RWD)". In other words, the upper atmosphere brightness temperature is considered to depend on the relative wind direction. The microwaves observed by the microwave radiometer mounted

on the satellite can be roughly divided into three elements as shown in Fig. 1.

$TB_{air\ emission}$ is microwaves emitted by the atmosphere, and TB_{ref} is microwaves emitted by the atmosphere scattered on the surface of the sea. $TB_{sea\ emission}$ is a microwave emitted from the sea surface.

The sum of each is the observed brightness temperature of the radiometer, which can be expressed by Eq. (1).

$$TB = TB_{air\ emission} + TB_{ref} + TB_{sea\ emission} \quad (1)$$

First, consider T . If the sea level is a calm mirror surface, T is expressed by Eq. (2).

$$TB_{0ref} = R_{0p} TB_{air} \quad (2)$$

R_{0p} is the reflectance, the subscript 0 is the mirror surface, and the subscript p is the polarization. Also, TB_{air} means downward atmospheric radiation. R_{0p} is expressed by Eq. (3) from Fresnel's reflection law.

$$R_{0p} = |\rho_p|^2, \quad \rho_p = \sqrt{\rho_v^2 + \rho_h^2} \quad (3)$$

$$\rho_v = \frac{\varepsilon \cos \theta_i - \sqrt{\varepsilon - \sin^2 \theta_i}}{\varepsilon \cos \theta_i + \sqrt{\varepsilon - \sin^2 \theta_i}} \quad (4)$$

$$\rho_h = \frac{\varepsilon \cos \theta_i - \sqrt{\varepsilon - \sin^2 \theta_i}}{\varepsilon \cos \theta_i + \sqrt{\varepsilon - \sin^2 \theta_i}} \quad (5)$$

where ρ is the reflectance coefficient of Fresnel, the subscripts v and h are vertically polarized waves and horizontally polarized waves, and ε is the complex permittivity, and θ_i is the angle of incidence. Also, as shown in Fig. 2, the angle of incidence and the angle of reflection are equal. Furthermore, since the emissivity of the sea surface is obtained by subtracting the reflectance from 1, $TB_{sea\ emission}$ can be formulated in the same manner.

The sum of these and the radiation from the atmosphere is the brightness temperature of the satellite-mounted microwave radiometer. Assuming that the observation dip of the microwave radiometer is the observation azimuth, this microwave radiometer can detect microwave radiation traveling in the direction of the zenith and azimuth. Although it is necessary to consider the beam width when actually calculating, the beam width is not considered because this chapter only explains the concept.

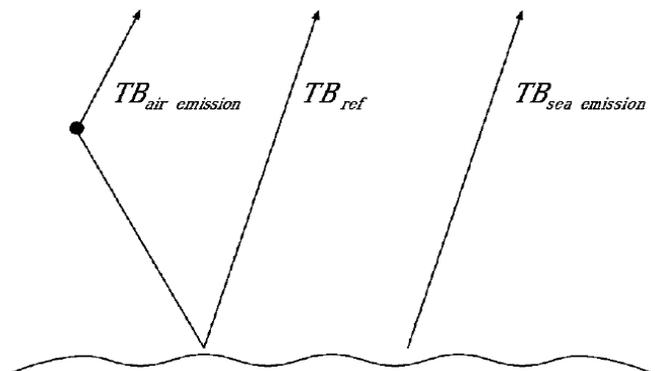


Fig. 1. Observed Brightness Temperature.

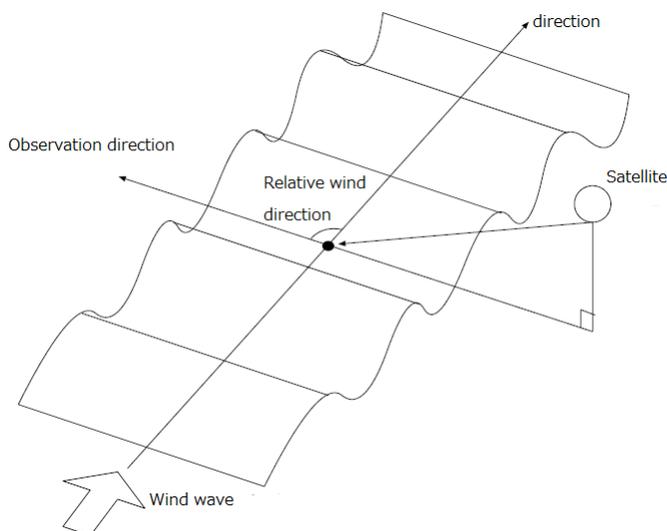


Fig. 2. Relative Wind Direction.

The observed brightness temperature TB_{ref} value of the microwave radiometer is the component of the zenith angle that scatters on the sea surface at the azimuth angle. Therefore, only the components of the zenith angle scattered in the azimuth angle are considered. In this paper, the angle of incidence in this case is called the observed angle of incidence. The actual sea level creates a complex gradient distribution due to various causes such as wind waves and swells.

Due to such sea level gradients, downward atmospheric radiation is scattered in all directions. It is said that the gradient distribution of the sea surface can be approximated by the Gaussian distribution, but strictly speaking, it is considered that the gradient distribution is distorted in the wind direction by the wind. That is, the wave created by the wind becomes a wave that travels in parallel with the wind direction, and the gradient distribution is biased. Although it is difficult to know the specific degree of bias, it is easy to imagine that winds and wind waves have characteristics that depend on the observation azimuth and the angle of the wind direction, as shown in Fig. 2. That is, the variance of the gradient distribution parallel to the wind direction is large, and the variance of the gradient distribution perpendicular to the wind direction is small. Therefore, the average observer firing angle differs between the case where the observation azimuth is parallel to the wind direction and the case where the observation azimuth is perpendicular to the wind direction.

When considering the local wave gradient, it is common to divide the wave into the smallest surfaces. By doing so, the reflection of the rough sea surface can be thought of as a specular reflection as shown in Fig. 3. In this paper, the average observation angle of the observation area at this time is set to the wind direction with the azimuth angle (RWD=0 deg. and / or 180 deg.), as shown in Fig. 4.

Temporarily, $\theta_{i1}=x^\circ$, then, $\theta_{i2}>x^\circ$, $\theta_{i3}<x^\circ$, $\theta_{i4}=x^\circ$. That is, $\theta_m=x^\circ$. In this case, there is also the influence of the rotation of the plane of polarization. Therefore, the following discussion advances to the front bank that the influence of the

rotation of the plane of polarization is small. Next, when the observation azimuth is perpendicular to the wind direction (RWD = 90 deg.), the observation incident angle changes as shown in Fig. 5.

Therefore, it is considered that the average incident angle increases as the relative wind direction approaches the vertical. As can be seen from equations (3) to (5), the reflectance changes as the incident angle changes. In addition, the state of change in reflectance differs between vertically polarized waves and horizontally polarized waves. Fig. 6 shows an example of the change in reflectance when the angle of incidence changes.

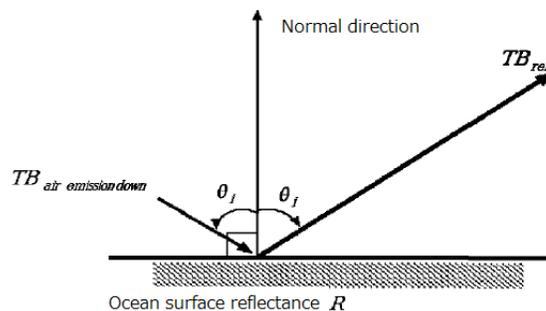


Fig. 3. Specular Reflection.

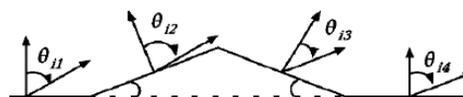


Fig. 4. Incidence Angle when Microwave Radiometer Observes in Parallel with Wind Direction.

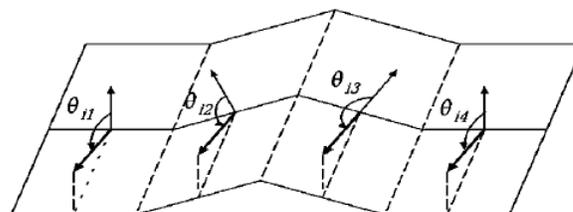


Fig. 5. Incidence Angle when Microwave Radiometer Observes in Perpendicular with Wind Direction.

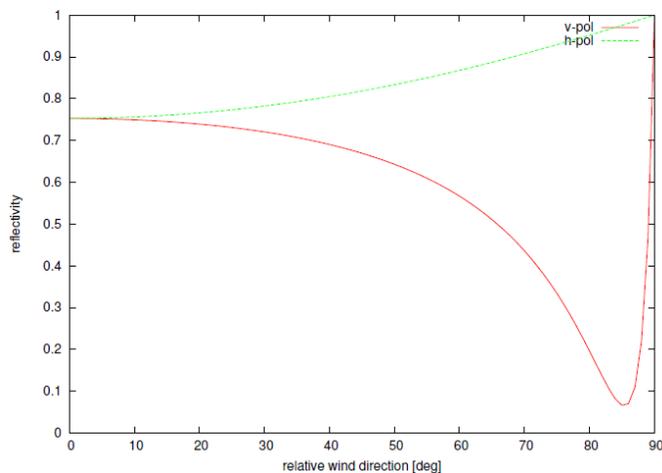


Fig. 6. Relation between Reflectivity and Relative Wind Direction.

First, in the case of horizontally polarized waves, the incident angle of ASMR-E is 55 deg. In the vicinity, the reflectance increases as the incident angle increases. On the contrary, reflectance becomes greater in accordance with the incident angle becomes smaller.

A similar relationship can be seen at 10GHz, 18GHz and 23GHz. Therefore, $RWD = 90 \text{ deg.}$, which increases the average observation angle of incidence. Then, the horizontal polarization component of the upper atmosphere brightness temperature becomes large, and the vertical polarization becomes small. In addition, $RWD \approx 0^\circ \text{ and/or } 180^\circ$, which has a relatively small average observation angle. Then, it can be expected that the horizontal polarization of the upper atmosphere brightness temperature becomes smaller, and the vertical polarization becomes larger. However, since the exact gradient distribution is unknown, it is not known how much the relative wind direction affects the brightness temperature. In this study, the effect was analytically derived.

IV. PROPOSED METHOD

In this study, the effect of the relative wind direction on the brightness temperature was analyzed by comparing the brightness temperature affected by the relative wind direction with the brightness temperature not affected. In this case, the brightness temperature includes not only the downward component of atmospheric radiation, but also the upward component of atmospheric radiation and radiation from the sea surface.

As the brightness temperature affected by the relative wind direction, the brightness temperature data (TB: Brightness Temperature) of Level 1B of the microwave radiometer AMSR-E mounted on the AQUA was used. AMSR-E Level 1B products have polarization types of horizontal polarization and vertical polarization, and the center frequencies of the frequency bands are 6.9, 10.7, 18.7, 23.8, 36.5, 89.0GHz, respectively. In this study, we used horizontal and vertical polarization in the frequency band below 23.8 GHz and used winter data that can be expected to have relatively strong winds and summer data that can be expected to have relatively weak winds.

Table I shows the time zones of the observation data. In the table, start time and end time indicate the observation start and end times in GMT, respectively. The observation positions are shown in Fig. 7. ASMR-E data for half a lap is 196 columns x about 2000 rows. It is the brightness temperature data of about 3920 opening points. Furthermore, the latitude range is relatively strong at latitude 60 north, from latitude 60 south. Therefore, we targeted about 261333 predictable data. Estimates at all points are shown in Table I.

TABLE I. TIME ZONES OF THE OBSERVATION DATA

No	Date	Start	End
1	Aug.11	23:21	00:11
2	Aug.12	11:43	12:33
3	Dec.8	23:27	00:18
4	Dec.9	11:49	12:39

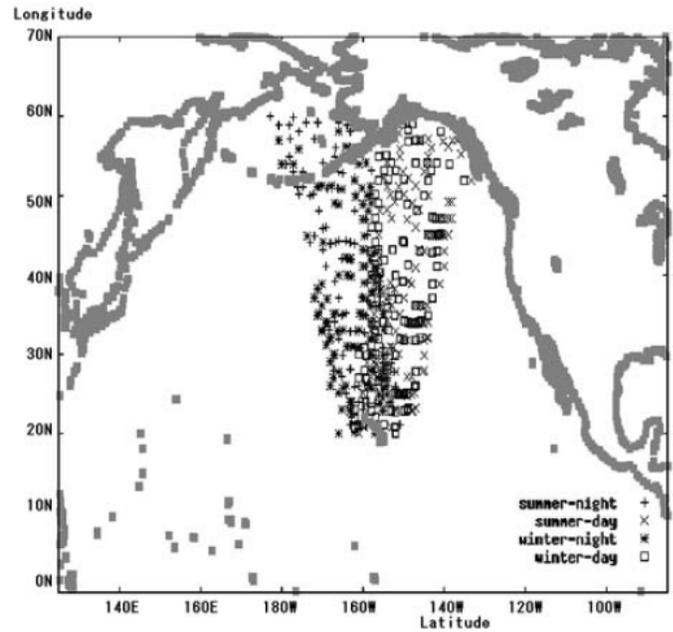


Fig. 7. Observation Positions.

Since it takes too much time, we used one-thousandth of the data and the data for GMT at random sampling. Therefore, the amount of data of the observed positions used is 261. We prepared representative summer / winter and day / night datasets and used them for the experiment. Wentz's microwave radiation transfer model based on global meteorological data (GTAD83.2: Global Tropospheric Analyses dS083.2) provided by the National Centers for Environmental Prediction (NCEP) as a brightness temperature that is not affected by the wind direction.

The brightness temperature derived using the above was used. The global meteorological data used at this time is the data at the same position and time as the ASMR-E data to be compared. The variables given as the input value of the microwave radiation transmission model are the absolute azimuth of the wind (WD: Wind Direction), the sea surface temperature (SST: Sea Surface Temperature), Precipitable water (PW), cloud water amount (CW: Cloud Liquid water). The output value is represented by the brightness temperature TB of the microwave. All these variables are obtained from GTAD83.2. GTAD83.2 has a mesh size of 1 by 1 degrees and is updated every 6 hours for ASMR-E data and GTAD.

Since the data of GTAD83.2 has a difference in the observation position, the data of GTAD83.2 was linearly interpolated to correspond to the data position of ASMR-E. The relative wind direction was defined by equations (6) to (8).

$$\theta_{amsr} = \begin{bmatrix} \cos(\theta_{amsr}) \\ \sin(\theta_{amsr}) \end{bmatrix} \quad (6)$$

$$\theta_{wind} = \begin{bmatrix} \cos(\theta_{wind}) \\ \sin(\theta_{wind}) \end{bmatrix} \quad (7)$$

$$RWD = \cos^{-1} \left(\frac{\theta_{amsr} \theta_{wind}}{|\theta_{amsr}| |\theta_{wind}|} \right) \quad (8)$$

RWD represents the relative wind direction, θ_{amsr} is the ASMR-E observation azimuth, and θ_{wind} is the wind azimuth.

The calculation result of the radiation transfer model is used as the data that is not affected by the relative wind direction, and it is defined by the following formula.

$$TB_{simu} = SIMU(sst, ws, pw, cw) \quad (9)$$

where TB is the brightness temperature that is not affected by the relative wind direction, $SIMU()$ is the microwave radiation transmission model, and the arguments are the sea surface temperature, sea wind speed, precipitable water, and cloud water obtained from GTAD083.2, respectively. Furthermore, the brightness temperature affected by the relative wind direction is represented by TB_{amsr} , and the relative influence of the wind direction is defined by the following formula.

$$\Delta TB = TB_{amsr} - TB_{simu} \quad (10)$$

Furthermore, since it is considered that the influence of the wind direction depends on the wind speed, the ΔTB call at a certain wind speed is represented by $\Delta TB/ws$. This relationship was approximated by Eq. (11).

$$\Delta TB|_{ws} \cong A RWD^2 + B RWD + C \quad (11)$$

Fig. 8 to 11 show examples of horizontal polarization in the 10 GHz band. Examples of band vertical polarization are shown in Fig. 12 to 13. From Fig. 8 to Fig. 13, it can be seen that the coefficients a_1 and c_3 in Eq. (11) change depending on the wind speed. The relationship was approximated by equations (12), (13), and (14).

$$A \cong a_1 WS^2 + a_2 WS + a_3 \quad (12)$$

$$B \cong b_1 WS^2 + b_2 WS + b_3 \quad (13)$$

$$C \cong c_1 WS^2 + c_2 WS + c_3 \quad (14)$$

Equations (11) to (14) and nine approximation coefficients from a_1 to c_3 can be used to add the RWD effect to the physical model. As a result, the coefficients in Table II were derived.

Examples of horizontal polarization in the 10 GHz band are shown in Fig. 14, 15, and 16.

From Fig. 8 to Fig. 10, the characteristics of the relative wind direction dependence become clear as the wind speed increases. That is, the observed brightness temperature increases as the relative wind direction approaches 90 degrees. This feature is prominent in horizontal polarization.

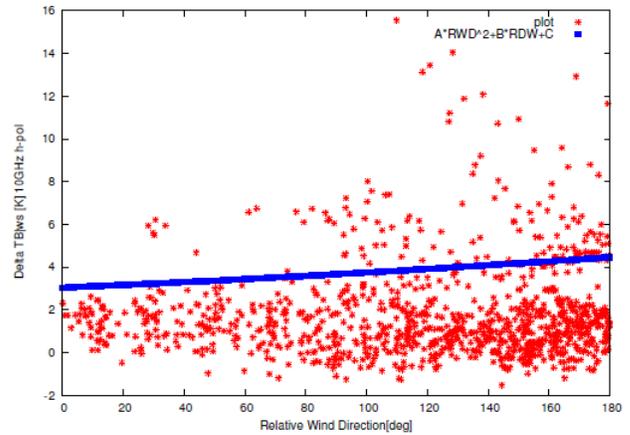


Fig. 8. Relation between Relative Wind Direction and Delta Brightness Temperature (10GHz H-pol 1.26 m/s).

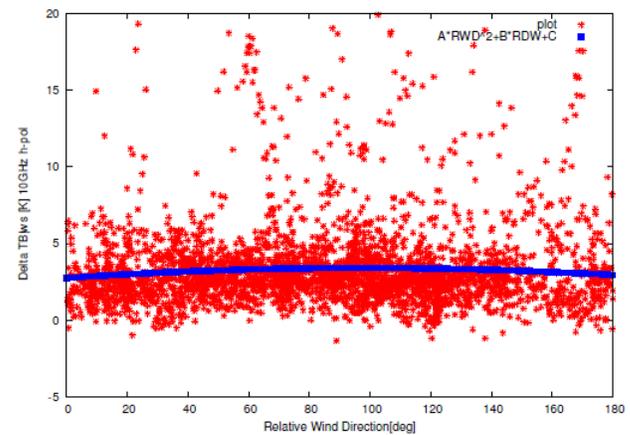


Fig. 9. Relation between Relative Wind Direction and Delta Brightness Temperature (10GHz H-pol 7 m/s).

TABLE II. COEFFICIENTS FOR EQUATIONS (11) TO (14)

		a1	a2	a3	b1	b2	b3	c1	c2	c3
6GHz	V-pol	0.011	-0.202	0.839	0.001	0.033	-0.702	0.007	-0.264	3.502
	H-pol	-0.003	-0.085	0.488	0.038	-0.251	0.42	-0.041	0.231	3.134
10GHz	V-pol	0.003	-0.071	0.0327	-0.002	0.033	-0.462	0.011	-0.253	3.518
	H-pol	-0.006	-0.036	0.303	0	0.448	-2.22	-0.027	0.115	3.26
18GHz	V-pol	0.005	-0.107	0.547	-0.003	0.0023	-0.684	0.014	-0.182	4.688
	H-pol	-0.006	-0.056	0.274	0.023	0.056	-0.383	-0.011	-0.049	6.286
23GHz	V-pol	0.01	-0.18	0.739	0.002	-0.187	0.719	0.025	-0.178	4.209
	H-pol	-0.004	-0.03	0.093	0.016	-0.097	0.761	-0.024	0.477	4.392

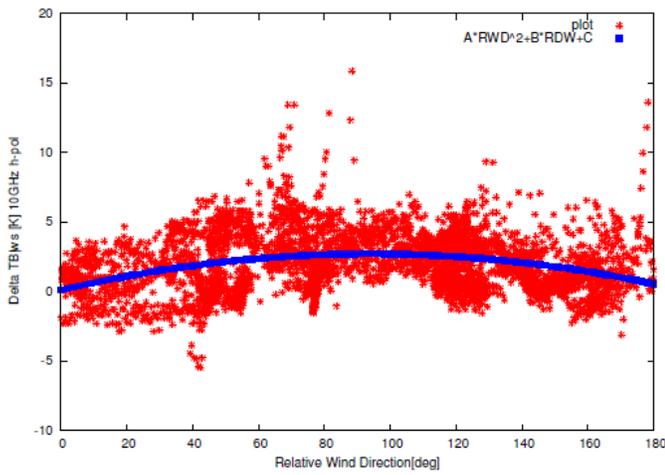


Fig. 10. Relation between Relative Wind Direction and Delta Brightness Temperature (10GHz H-pol 12.93 m/s).

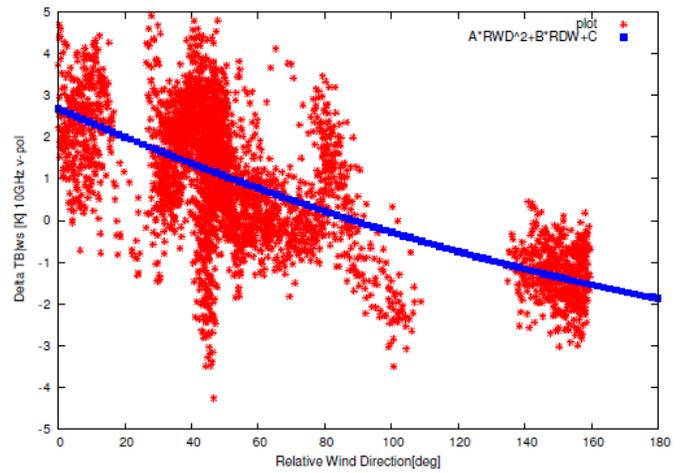


Fig. 13. Relation between Relative Wind Direction and Delta Brightness Temperature (10GHz V-pol 19.09 m/s).

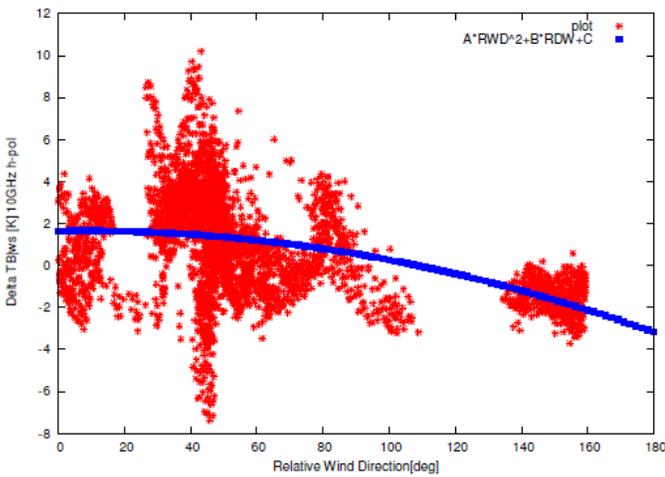


Fig. 11. Relation between Relative Wind Direction and Delta Brightness Temperature (10GHz H-pol 19.09 m/s).

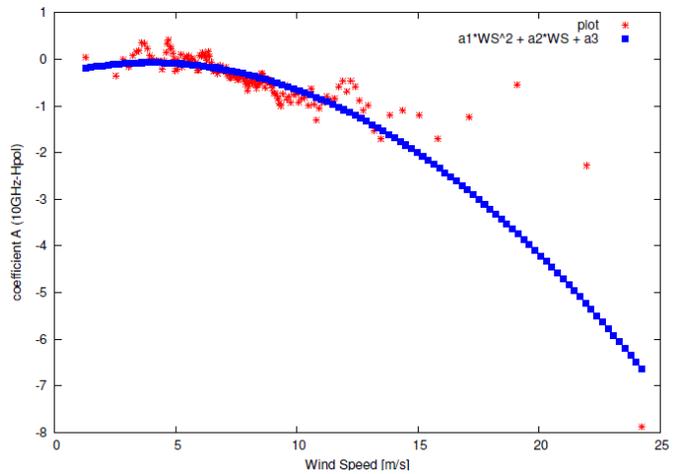


Fig. 14. Relation between Coefficient A and Wind Speed (10GHz H-pol).

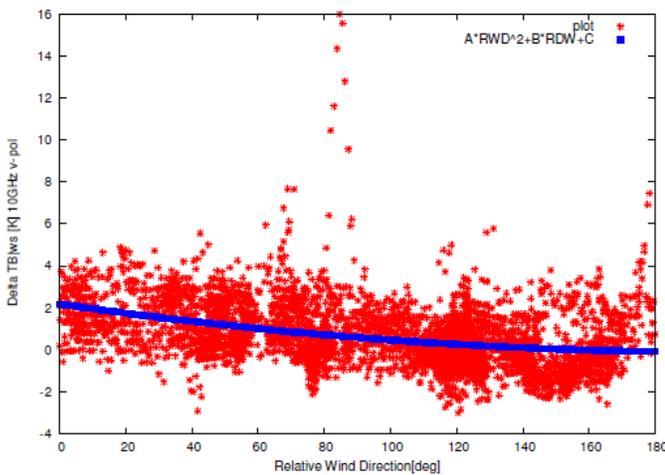


Fig. 12. Relation between Relative Wind Direction and Delta Brightness Temperature (10GHz V-pol 12/93 m/s).

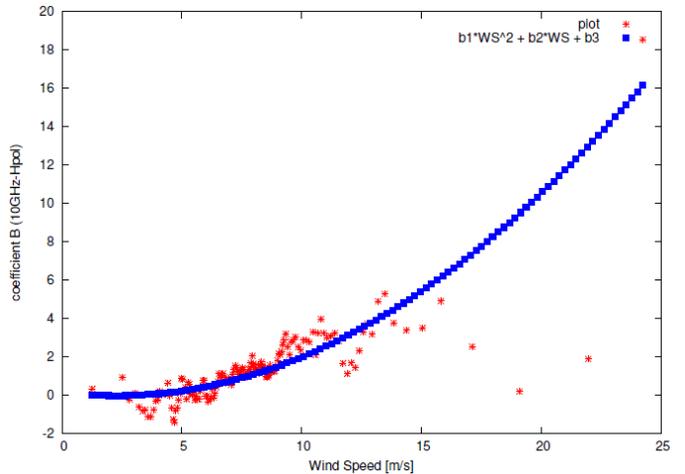


Fig. 15. Relation between Coefficient B and Wind Speed (10GHz H-pol).

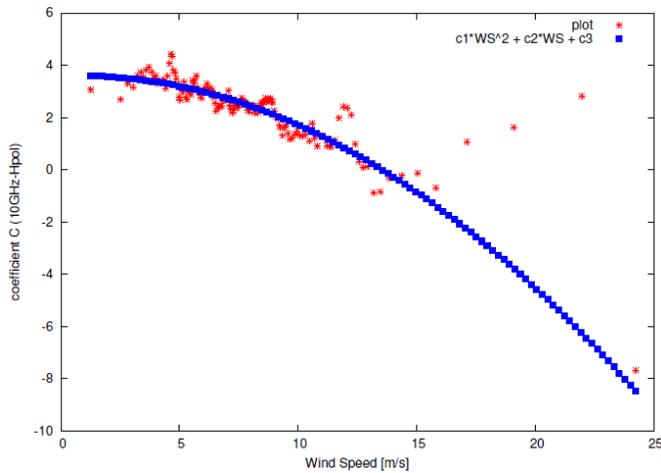


Fig. 16. Relation between Coefficient C and Wind Speed (10GHz H-pol).

As shown in Fig. 14 to Fig. 16, the change in ΔTB when the wind speed exceeds about 15 m/s is clearly different from that when the wind speed is below 15 m/s. It is known that wind waves grow when the wind speed increases and collapse when the wind speed exceeds a certain level. From this, it is considered that the feature of about 15 m/s is due to the collapsing waves. Another possible cause is the white waves that occur when the wind speed is strong. In addition, Fig. 12 and Fig. 14 clearly show the difference of ΔTB when observed from the front and the rear of the collapsing wave.

When the RWD is 90 degrees, that is, when observing from behind the collapsing wave, the observed TB becomes large; when $RWD > 90$ degrees, that is, when observing from the front of the collapsing wave, the observation becomes small.

V. EXPERIMENT

Sea wind direction, wind speed, sea surface temperature, precipitable water, cloud water estimations are conducted. Using Equations (11) to (14) and Table II described in the previous chapter, the effect of the relative wind direction on the brightness temperature can be understood. Therefore, by adding the brightness temperature derived using this equation and the coefficient to the microwave radiation transfer equation of Wentz, the radiation transfer equation considering the relative wind direction can be obtained.

This function is non-linear. Therefore, to determine the physical quantity at sea from the brightness temperature at the upper end of the atmosphere, the inverse problem must be solved. To solve this reverse problem, we used Simulated Annealing, which was proposed by Arai. The luminance temperature used for the estimation was ASMR-E Level 1B (horizontal and vertical polarization 6,10,18,23 GHz band), which was the same as the data used in the previous chapter. The estimated physical quantities are sea surface temperature SST, sea surface temperature WS, relative wind direction RWD, precipitable water PW, and cloud water volume CW, and the estimated location is over the Pacific Ocean (20-60 north latitude, 180-230 east longitude).

The correct answer data used to compare the estimation results is the global meteorological data GTAD83.2 provided by UCEP, which is also the same as the one used in the previous chapter. As in the previous chapter, linear interpolation was performed to positionally correspond to the ASMR-E data. The standard deviation between the value estimated from the luminance temperature of ASMR-E and the correct answer value obtained from GTAD83.2 was used as the standard deviation. The standard difference was derived by Equation 15.

$$Acu = \sqrt{\frac{\sum_{i=1}^n (Ans_i - X_i)^2}{n}} \quad (15)$$

where Acu is the estimation accuracy, Ans_i is the correct value obtained from GTAD83.2, X_i is the estimated physical quantity, n is the data number, and d is the total number of estimated data. In addition, when calculating this estimation accuracy, the estimation point where the error (energy E) due to simulated annealing was clearly large was rejected because the estimation result is clearly wrong. Table III summarizes the estimation accuracy of each estimated physical quantity. In addition, Table III also summarizes the estimation accuracy when estimating with the conventional model as a target for comparison. The data numbers are shown in Table I.

The conventional model is a method using Wentz's microwave radiation transfer equation and improved simulated annealing by Arai and Sakakibara. Hereafter, the algorithm when considering the influence of the relative wind direction derived in this study is called the modified model and is distinguished from the conventional model. The modified model is Wentz's microwave radiation transfer equation modified by the influence of the relative wind direction shown in the previous chapter. Regarding the wind speed, the estimation accuracy improved except for data 1. The reason why the estimation accuracy of data 1 deteriorated is that it is estimated at the same time.

It is probable that the estimation accuracy of the relative wind direction was poor. On the contrary, the data 3 in which the estimation accuracy of the relative wind direction is greatly improved also greatly improves the estimation accuracy of the wind speed. It can be seen that to improve the accuracy of wind speed estimation in consideration of the relative wind direction, it is necessary to improve the estimation accuracy of the relative wind direction. However, the effect of the relative wind direction on the observed luminance temperature is small compared to other physical quantities estimated at the same time. Therefore, in the estimation method used in this study, it is considered that the estimation accuracy of the relative wind direction deteriorated due to the error that occurred when calculating other physical quantities.

Compared to data 1 and data 2 observed in summer, the estimation accuracy of wind speeds in data 3 and data 4 observed in winter has improved significantly. This is thought to be because the relative wind direction has a stronger effect in winter when the wind is strong than in summer when the wind is weak.

TABLE III. SUMMARY OF THE ESTIMATION ACCURACY WHEN ESTIMATING WITH THE CONVENTIONAL MODEL AS A TARGET FOR COMPARISON

Model	No.	SST[K]	WS[m/s]	RWD[deg]	PW[kg/m ²]	CW[kg/m ²]
Existing	1	2.63	2.13	86.92	2.45	0.004
	2	1.26	2.39	67.69	1.28	0.08
	3	1.53	2.01	62.65	2.92	0.08
	4	1.32	2.52	52.64	2.75	0.07
Proposed	1	0.96	2.15	99.76	2.25	0.04
	2	1.19	1.91	56.53	1.83	0.1
	3	1.71	0.78	18.12	1.3	0.01
	4	1.37	1.76	68.02	1.29	0.05

Since the estimation method used in this study is a method of estimating physical quantities at the same time, the estimation accuracy of other physical quantities is affected as well as the change in the estimation accuracy of the wind speed, as shown in Table III. The accuracy of sea surface temperature estimation has improved in summer, and the accuracy of precipitable water estimation has improved in winter.

VI. CONCLUSION

By comparing the observed brightness temperature affected by the relative wind direction with the observed brightness temperature not affected, the effect of the relative wind direction on the observed brightness temperature was found. In particular, we were able to extract clear features in horizontally polarized waves. In addition, these features could be expressed by equations and coefficients. In the case of the data used in this study, it was found that the estimation accuracy of the wind speed changed by using the model in which the proposed relative winds direction affects the luminance temperature.

When the estimation accuracy of the relative wind direction estimated at the same time was relatively good, the estimation accuracy of the wind speed was also improved. However, if the estimation accuracy of the relative wind direction estimated at the same time is poor, the estimation accuracy of the wind speed may also deteriorate. Therefore, it was also found that it is necessary to improve the estimation accuracy of the relative wind direction in order to improve the estimation accuracy of the wind speed by the proposed model. In addition, the proposed model is a symmetric function with the observed brightness temperature deviation as the peak when the relative wind direction is 90 degrees, and it is considered that the deviation becomes smaller when the relative wind direction is smaller or larger than this.

Considering that the quadratic function approximation is the first-order approximation, the relationship between the relative wind direction and the observed brightness temperature was approximated by a quadratic equation. However, it is necessary to consider a more appropriate approximation function in consideration of physical grounds.

In this study, the effect of the relative wind direction on the luminance temperature was attributed to changes in reflectance. Especially when the wind speed exceeds about 15 m, the optimum approximation formula can be obtained by understanding the physical phenomena related to the relative wind direction, such as the change in brightness temperature due to the collapsing waves that occur when the wind speed is strong, it is conceivable then.

In this study, the proposed method was evaluated by comparing it with the conventional method using NCEP GDAS as the correct answer, but the errors included in NCEP GDAS will be examined in the future. The error analysis and sensitivity solution in the simultaneous estimation of multiple physical parameters and the simultaneous estimation for improving the accuracy of wind direction determination are left to the references.

VII. FUTURE RESEARCH WORK

In this paper, we have shown a method for classifying each pixel for a certain image, but the proposed method can be applied not only to images but also to various sets with values for each element, such as remote sensing, GIS, and it is considered that it can be widely applied to other data mining. Further research works are required for the other applications not only rice paddy field detection with SAR imagery data.

ACKNOWLEDGMENT

The authors have received regular discussions, useful opinions and suggestions. Remote Sensing Systems Co., Ltd., Dr. Frank J. Wentz, Professor Masanori Neda of Kyoto University, Professor Naoto Ebuchi of Hokkaido University, JAXA Akira Shibata I am deeply grateful to the doctor. In addition, the master's degree of our graduate school that cooperated with the data collection experiment and analysis.

We would like to thank Mr. Atsushi Sakakibara of former student of Saga University and Prof. Dr. Hiroshi Okumura and Prof. Dr. Osamu Fukuda of Saga University for his valuable comments and suggestions.

REFERENCES

- [1] Kohei Arai: Basic Theory of Remote Sensing, Academic Book Publisher, 2001.
- [2] Kohei Arai: Self-study Remote Sensing, Morikita Publishing, 2004.
- [3] Japan Aerospace Exploration Agency: -Handbook for Utilizing Earth Observation Data — AMSR-E Flat — ”, 2006. http://www.eorc.jaxa.jp/hatoyama/amr-e/amr-e_handbook "(6Jun. 2007).
- [4] Frank Wentz: ASMR Ocean Algorithm, Remote Sensing System, 2002.
- [5] Akira Shibata: Observation of atmosphere, ocean, and land with multi-wavelength microwave radiometer, Journal of Japan Remote Sensing Society Special Feature: Radio Waves and Remote Sensing, 12,1,59-64, 1992.
- [6] Arai, K. and J. Sakakibara: Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Advances in Space Research, 37,12, 2202-2207, 2006.
- [7] M. Konda, A. Shibata, N. Ebuchi and K. Arai: Correction of the effect of relative wind direction on wind by AMSR, Journal of Oceanography, 64,395-404, 2006.
- [8] Dennis J. Shea: "An Introduction to Atmospheric and Oceanographic Data", 1994. <http://dss.ucar.edu/docs/data-intro-technote/tn-404.pdf>

- [9] M.Matsumoto and Kohei Arai, Simplified expression of the radiative transfer equation in thermal infrared windiow spectrum, Proc.of the IGARSS'93, 673-675, 1993.
- [10] N.Ebuchi, Kohei Arai, et.al., Evaluation of vector winds observed by NSCAT in the seas around Japan, Journal of Ocean Society of Japan, Vol.56, No.5, pp.495-505,(2000).
- [11] Kohei Arai, Polarization sensitivity of the ocean surface together with wind vector derived from POLDER and NSCAT on ADEOS, Proceedings of the NASA Oceanography Scientific Conference, Florida, USA, 2001.
- [12] Kohei Arai and Jun Sakakibara, Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Abstracts of the 35th Congress of the Committee on Space Research of the ICSU, A1.1-0130-04, (2004).
- [13] Kohei Arai and Jun Sakakibara, Simultaneous estimation of sea surface temperature, wind speed and water vapor with AMSR-E data based on improved simukated anneiling, Proceedings of the Renewable Energy Resources Symposium, 00547, 2006.
- [14] Kohei Arai, Space and time retrieval of tide wind speed and wave height with altimeters onboard satellites based on PostGIS system, Proceedings of the Renewable Energy Resources Symposium, 00548, 2006.
- [15] Kohei Arai and J.Sakakibara, Estimation of SST, wind speed and water vapor with microwave radiometer data based on simulated annealing, Advances in Space Research, 37, 12, 2202-2207, 2006.
- [16] M.Konda, A.Shibata, N.Ebuchi and Kohei Arai, Correction of the effect of relative wind direction on wind speed derived by AMSR, Journal of Oceanography, 64, 395-404, 2006.
- [17] Kohei Arai, Data fusion between microwave and thermal infrared radiometer data and its application to skin sea surface temperature, wind speed and salinity retrievals, International Journal of Advanced Computer Science and Applications, 4, 2, 239-244, 2013.
- [18] Kohei Arai, Comparative Study of optimization Methods for Estimation of Sea Surface Temperature and Ocean Wind wit microwave Radiometer data, International Journal of Advanced Research on Artificial Intelligence, 5, 1, 1-6, 2016.

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 a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 55 books and published 620 journal papers as well as 450 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. <http://teagis.ip.is.saga-u.ac.jp/index.html>

A Study of Security Impacts and Cryptographic Techniques in Cloud-based e-Learning Technologies

Lavanya-Nehan Degambur, Sheeba Armoogum, Sameerchand Pudaruth
ICT Department, University of Mauritius, Moka, Mauritius

Abstract—e-Learning has transposed the perception of teaching and learning considering knowledge delivery and knowledge acquirement. Today, e-learning participants access and upload their materials at any time and at any place since e-learning technologies are typically hosted on the cloud. Cloud computing has embellished the base platform for the future of e-learning, however, security and privacy remains a major concern. Cloud-hosted e-learning technologies as they are accessed over the internet suffer from the same risks to information security aspects namely availability, confidentiality, and integrity. In such a context, data authenticity, privacy, access rights and digital footprints are vulnerable in the cloud. Research in this domain focuses on specific components of cloud and e-learning without covering a holistic view of applied cryptographic techniques and practical implementation. Hence, aiming at the various security aspects and impacts of cloud-based e-learning technologies, this paper puts forward reviewing the various cryptographic techniques used to secure data across the whole end-to-end cloud-based e-learning service spectrum using systematic review and exploratory method. The results obtained define several sets of criteria to evaluate the requirements of cryptographic techniques and propose an implementation framework across an end-to-end cloud-based e-learning architecture using multi-agent software.

Keywords—e-Learning; cloud computing; data management; pseudonymization; data deduplication

I. INTRODUCTION

In a world where the internet is the most potent communications enabler, humanity has transcended orthodox teaching and learning ways and made them global with e-learning where information, communication and digital technologies are utilized to ease the learning process. The e-learning concept has offered an effective educational tool that is accessible from anywhere to anyone encompassing professionals, scholars, teachers, and students by combining the virtues of the internet and the wisdom of knowledge. The efficacy of e-learning is improvised by enhancing the training of teachers, curriculum developments, assessments reforms and infrastructure optimization in a holistic manner. e-Learning has enriched the economic growth in various countries and has reduced the digital divide between countries, societies, and communities. The introduction of e-learning technologies in patriarchal or male-supremacist communities allows the emancipation of girls without bypassing societal norms. Underserved students can make use of this method to study at their own pace and improve their participation abilities and cognitive growth. Hence, through the dissemination of programs, gender gaps can be narrowed across the whole human diaspora. The recent COVID-19 pandemic has

increased the demand for e-learning platforms since traveling has been inhibited owing to recurrent lockdowns. Thus, to be fully transformative, e-learning should be integrated formally into curriculum creation and teacher training and informally into the habits of students [1]. Cloud computing platforms have diversified the conceptualisation of delivering education using modern e-learning methods. However, this base platform for the future of e-learning is vulnerable to security threats and privacy issues.

A. Modern e-Learning Technologies

Modern e-learning involves the delivery of courses to people in an automated or virtualized form such as videos and interactive methods or via formal teaching from teachers in a personal form via the internet using digital technologies [1, 2]. Virtual learning leans on the visual acumen of protagonists to help the users present and understand topics, study at any time anywhere without teacher intervention, the ability to edit, update and share materials without prior notification and synchronization between teachers and students. Moreover, it can provide an increased number of courses without logistical investment; increase the number of students owing to the flexibility, and cost-effectiveness of such learning. Using the virtualized form, or virtual learning, students and teachers can share a variety of resources, topics and materials and present them in a customizable and personalized format to ease teaching and comprehension. Whereas in a personal form or personal learning, students and teachers have a personal space on a single-use instance or continuous use instance and use their materials to learn and improve skills and teach respectively. Features of personal learning are the ability of users to manage teaching sessions and learning materials in learning platforms, multi-user interaction during e-learning sessions and performance and goal setting per user profile.

Modern e-learning is cloud-based whereby the e-learning platforms are hosted on the cloud instead of hardware and software being installed, run, and administered on the learning provider's premises. The educational materials in today's e-learning are virtualized in cloud infrastructures and it is up to the cloud service providers to guarantee the uptime of the services available to e-learning protagonists. Cloud-based e-learning platform has triumphed due to the abilities of remote access, cost efficiency, open research-oriented environments, ability to analyze and provide insights about the behaviors of protagonists, the disintegration of geographical barriers and time constraints.

However, where information is present, dangers to information are omnipresent and information traveling on the

internet is constantly exposed to security threats. Since e-learning systems are diversified, there is a variety of resources and consumers of those online resources. Collaboration, interconnectivity, and information sharing are the pillars of e-learning systems so that data and the information it induces must be protected to maintain confidentiality, integrity, and availability. Information security issues in e-learning include data manipulation, confidentiality compromises and fraudulent authentications as e-learning developments always lean towards interoperability between learning environments, heterogeneous devices, multi-technology applications, and academic discoveries. Cloud computing readily provides this unified interconnected requirement for successful e-learning technologies.

B. Cloud-Computing and e-Learning

One exclusively pays the cloud services to fit and enhance business needs, administer infrastructure optimally and reduce operational costs. A wide-scale pooling of computing and networking resources such as processing, memory, storage, and bandwidth, all available on-demand, achieves this. Accessing those resources requires flexibility because resource distribution needs to be precise and fast to cater to rapid fluctuations in demand without service disruption or quality deterioration [2, 3]. Cloud-based e-learning platforms use cloud computing service models such as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) that are deployed in three deployment models named public cloud, private cloud and hybrid cloud. IaaS provides the physical IT infrastructure and architecture of the system that the clients will use and allows the clients to control only the infrastructure resources provided to them, not the underlying cloud infrastructure hosting the IaaS.

PaaS provides services in terms of the operating system, hosting software and application development lifecycle-software using which the customers can develop their applications that are run on virtual machines, which abstracts the platform from the underlying physical infrastructure. SaaS offers software applications as services over the internet as compared to usual software packages bought by individual clients [4]. The end-user can hence use the cloud service provider's applications hosted on the cloud from anywhere at any time without catering for the management and control on the underlying platform or infrastructure. The public cloud deployment model delivers the cloud services readily available to the public while private cloud services by an organization are made available only to that organization and selected protagonists. Hybrid cloud computing is a mix of public and private cloud deployments to create an automated, unified, and fault-tolerant environment that offers modulated services based on user usage and requirements.

Accordingly, cloud-hosted e-learning technologies depend on the internet-connected cloud resources to function. However, both the cloud-based e-learning technology and e-learner are under constant threats by being connected with the internet. The COVID-19 pandemic has induced a rapid growth in cloud-based e-learning usage but this has also amplified attacks on such technologies. The main security concerns to cloud computing are browser security, authentication, privacy, duplication, and availability while those of e-learning are user

authentication and authorization, data confidentiality and blocking, denial of service and flooding attacks. Thus, the similitude between security concerns of cloud computing and e-learning can be easily observed [2, 3].

This study aims to conceptualize how e-learning services are provided by cloud technologies, discover the various cyber security issues that impact cloud-based e-learning technologies, study the state of art cryptographic techniques used to address the issues in several aspects of cloud-based e-learning technologies, find discrepancies in application and implementation of such cryptographic techniques and finally propose solutions to those problems. The proposed approach will be beneficial for the research community to identify the set of criterias to evaluate different cryptographic techniques.

This paper is structured as follows: Section 2 relates the literature review on the security impacts upon the usage of e-learning and cloud computing technologies. Section 3 describes the research methodology used to perform the study. Section 4 analyses the findings and provides results and Section 5 concludes the paper.

II. LITERATURE REVIEW

In this section, a detailed review is provided on the security impacts on the usage of e-learning and cloud computing technologies and the different cryptographic techniques applied along with the entire end-to-end cloud-based e-learning service architecture.

A. e-Learning Technologies Security Aspects, Impacts and Threats

Security threats on e-learning are the problems that can adversely influence the safety of e-learning end users and their data and also apply to user authentication, authorization, and confidentiality. If the usernames and passwords given to users to log onto the e-learning platforms are compromised, users may lose the ability to use the e-learning platforms and may risk their personal, professional, and confidential transaction information being accessed and misused by unauthorized parties. Compromised systems may be vulnerable to blocking attacks whereby an attacker attacks a user's e-learning content and access e-learning material and flooding attacks whereby an attacker bombards the e-learning platform with bogus requests using an account so that the user loses access time [2]. In addition to user authentication, concerns are raised for user operation process tampering to damage data patterns and social behavior deduction through patterns to impede user privacy.

According to the study in [2, 5], Short Message Service texts (SMS) can be used as two-factor authentication to complement usernames and passwords and prevent unauthorized access; biometrics such as fingerprints, iris recognition, or voice recognition can be used with attribute-based cryptography; digital signatures can be used to authenticate identity and integrity of data and also non-repudiation of transactions; access control lists and processes can be applied to the server and user resources access to customize access mechanisms. SaaS security by default is used to secure e-learning applications. Hence, the whole system infrastructure of servers and storage must be considered in terms of security impacts. To reduce threats, the authors

recommended including server, disaster recovery, and safety and management aspects.

Aside from user authentication and authorization, private information protection, and data integrity protection; raising awareness, learning resources authenticity, seamless access, location privacy, digital rights, and usage anonymity should also be catered for e-learning security. Doing so will fulfill basic security criteria such as availability, integrity, confidentiality, authenticity, non-repudiation, and accuracy. Maitra and Bhatia [6] used vulnerability scanners such as Netsparker and N-Stalker to test e-learning platform vulnerabilities and proposed methodologies to ensure security. Both scanners identified the following vulnerabilities: SQL injection, cross-site scripting, directory traversal, BREACH attack, JavaScript library vulnerabilities, CRIME SSL/TLS attack, HTTP response splitting/CRLF injection, file inclusion, HTTP parameter pollution, HTTP authentication, and insecure cookies. According to Maitra and Bhatia [6], e-learning platform security can be accomplished in two proposed approaches: hierarchical and distributed. The hierarchical approach involves applying security in a top-down centralized manner across components while distributed approach applies different security models for every different component in the e-learning system while allowing interaction.

Although e-assessment, the use of integrated information technologies to support assessment processes across all stages in its life cycle in e-learning [7] has proliferated, it has faced three main threats: identity misuse, disclosure of information, and fraudulent alteration. During an e-assessment session, identity misuse may happen if an attacker uses the identity of the real e-learning user being assessed; sensitive and private data transmitted may lead to the disclosure of confidential information. Since e-assessment and e-learning data are stored in databases, they can be subject to alteration which is problematic to students and teachers in terms of data integrity. Those issues can be analyzed and tackled in two perspectives namely the educational perspective by considering learners and teachers scenarios and problems and using the technical architectural perspective to secure the information system running the e-assessment solutions [7]. According to [8], the e-learning system, as well as the underlying infrastructure, should be evaluated for threats and risks to secure the subsequent technological solutions analysis and security policy audits by financial institutions and payments.

Cloud-based m-learning extends the e-learning by expediting the delivery of educational activities, materials, and contents that can be readily performed and accessed at any time at any place via the internet using mobile devices. Since mobile devices are exposed to operating system vulnerabilities, mobile web browser vulnerabilities, untrusted applications, application collusion, spyware, malware attacks, data leaks, jailbreaks, and personal user malpractices; cloud-based m-learning is posed to vulnerabilities. These vulnerabilities can be analyzed in a three-tiered architecture comprising of mobile device tier, network platform/provider tier, and cloud tier. Malicious push advertisements and SMS's can be sent using mobile devices to perform distributed attacks. Vulnerabilities due to network tier threats are inherent to the service provider such as internet protocol vulnerabilities, Man-In-The-Middle

(MITM) attacks, and malicious server agents. Generally, the cloud platforms are vulnerable to DNS/web spoofing, in-house unauthorized access, malware injection attack, authentication, and MITM attack, side-channel attack, virtual machine escape, and Denial of Service (DoS) attacks [9]. Conventionally, cloud-hosted applications are affected via attacks like cross-site forgeries and SQL injections; viruses and e-learning modules from session hijackings and riding affect Learning Management Systems (LMS). Besides, cloud data storage is susceptible to obsolete and insecure cryptographic storage and data duplication.

B. Cloud Computing Security Aspects, Impacts and Threat

Security concerns related to cloud computing are associated with basic security on aspects of data such as transmission, storage, and recovery, availability of applications, services, and data authentication demands, and browser security [2]. From a customer perspective and owing to the abstract nature of cloud and lost physical control, the following five main security threats have been obtained: data exposure to unauthorized antagonists, unauthorized access, data loss, manipulation and induction, Service Level Agreement (SLA) violation and privacy breaches. Data exposure and unauthorized access affect the data confidentiality requirement; data loss and data manipulation breaches data integrity requirement; privacy breaches antagonizes privacy preservation requirement while SLA violation goes against both data confidentiality and integrity.

Threats faced by e-learning users and providers are mostly internet-based where the Denial-of-Service (DoS) and Distributed DoS (DDoS) are omnipresent threats to the availability of such services. Rahman et al. [5] subsequently explored the different attacks on cloud-based e-learning platforms. Firstly, the web browser, which is the user's gateway to the e-learning platform exposed an ideal target for an attack. Phishing attacks eventuate when during the authentication process the user access has not been verified and certified. However, backdoor channel attacks negate authentication to prevent trusted users from accessing their confidential data, and virtual machine attacks involve seizing vulnerabilities in virtualization platforms to detach the physical resources to the virtual resources and reroute data access towards hackers. Nevertheless, insider attacks involve people familiar with the e-learning service provider getting access to the system through knowledge of system policies and architectures. Third-party cloud providers outsource cloud resources among themselves to provide high availability; however, this can be detrimental to user data privacy if policies, procedures, and contracts are not clearly defined. To prevent theft, unauthorized induction, and dissemination, they suggest disposing of the imperfect and redundant data from the cloud and physically.

Since cloud computing is provided in three distinctively different service models, each of them has its own security requirements which should be analyzed. In SaaS, security impacts are about data security, network security, data integrity, data segregation, and data breaches. Security issues regarding PaaS affect data location, and privileged access while those in IaaS are web service attacks, SLA attacks, DDoS, MITM attacks, and DNS security. Data location

vulnerabilities arise because it is unknown where data is stored and processed physically when users are accessing the applications on the platform [10]. Whereas privileged access vulnerabilities occur as a consequence of cloud providers potentially having all possible access rights to data residing on their platforms and any breach on the cloud provider side cascades onto the data.

To use a cloud service over the internet, the cloud provider runs web service protocols that are exposed to attack using XML signatures breaking security between a user's browser and the cloud service. Since SLAs are the legal bindings to service delivery between providers and users, attacks are possible against SLAs to exploit undefined metrics and hence defy quality, availability, performance, and reliability of resources. DDoS attacks deny important services from running by unleashing a tremendous number of requests, which are difficult to be handled by the attacked service. Usually, a Master controls several Slave bots to launch a DDoS on the Target cloud IaaS server. MITM attacks, a subcomponent of eavesdropping, happen when the attacker positions itself between the cloud user and the cloud IaaS server to hear and retrieve communication and even falsify the connection [10]. Besides, DNS attacks occur when vulnerabilities in domain name services are exploited to prevent the resolution of the IaaS cloud environment's domain names to the correct IP addresses. Both the cloud users' and the providers' packets can then get rerouted to prevent access, provoke a DoS, compromise credentials, falsify connections or clone queries and requests.

C. Cryptographic Techniques in e-Learning Technologies

To secure e-assessment in terms of personal data protection and hosts and network protection, the TeSLA system proposed transport layer security (TLS) using authorization certificates, public key infrastructure, and pseudonymization. TLS allows every entity to mutually authenticate with its peers and create tunnels secured with data encryption and integrity checks using X.509 certificates instead of passwords. TeSLA's PKI manages certificates employing different certificate authorities (CA), revocation lists, three-layered security procedures, and 4096 bits RSA (Rivest-Shamir-Adleman) keys for Certification Authority (CA) certificates [7]. Identity management and data protection are achieved with pseudonymous credentials adapted from attribute-based signatures utilizing randomized TeSLA IDs generated per user so that the full identity of a user remains anonymous.

In contemplation of securing cloud-based m-learning architecture, each tier is secured in a top-down hierarchical manner starting with mobile devices with multi-client authentication, multiple firewalls, and network and exchange servers. Authentication and authorization protection, identity and key management, backup and disaster recovery, anti-replay techniques, state-of-the-art encryption, and protection as a service are solutions to protect the cloud infrastructure tier. Lastly, the network tier is protected using next-generation firewalls and application access authorization [9].

To secure and authenticate data storage in e-learning systems, hash tables and hash trees have been proposed to ensure data integrity for the transmission, storage, and

processing of authentication data between a user and the e-learning system. Encryption is performed on authentication information before transmission and a linear dimensional reduction transformation projects user data and verification data into lower dimensions to preserve relative vector distances and authentication correctness. To secure authentication data integrity, InterPlanetary File System (IPFS) is proposed to store data on a Merkle Directed Acyclic Graph (DAG) file structure, which combines a Merkle tree and a Guided Ring graph. IPFS uses SHA-1, SHA-256, and BLAKE2 cryptographic hash functions to guarantee immutability and immunity to DDoS attacks on the authentication process [11].

To protect multi-agent e-learning platforms' access control and ensure trust and reputation, a combination of Role-Based Access Control (RBAC) inspired models called Trust-Based Access Control (TrustBAC) and Trust Satisfaction and Reputation (T-SR) have been proposed by Asmaa and Najib [12]. Trust levels between users/actors/agents can usually be generated by using user credentials, results of past user interactions, user characteristics, and the context in which those actions occurred [12]. This allows conditions to be created to define safety rules to be applied in access control processes, simplify the generation of trust value, and trust establishment.

Cyber-trust provides legal significance to a public exchange of documents over the network in e-learning environments and hence helps counteract cyber-attacks and cyber-espionage. Network Time Synchronization (NTS) provides standardized cyber-trust assurance for e-learning systems by using three layers of trust criteria expanded to the public internet cyberspace together with an independent Network Time Synchronization source. The three layers of trust criteria are the basic factors that combine to granularly model what a legal user should be and are based on five trust characteristics targeted, subjective, measurable, dynamics, and conditionally passed [13]. Independent Network Time Synchronization source is used because in-built time synchronization modules and synchronization subnets based on network time protocol can be compromised which can threaten end-user and e-learning platform private keys and digital signatures while also endangering the whole public key infrastructure used to secure the communication.

Ali and Zafar [14] recommended that, for e-learning security, DoS attacks and unauthorized logins can be handled by single sign-on authentication. Data evaluation issues at login and on course contents are solved using trust certificates and biometrics. Architecture challenges concerning data transmission channels and access controls are catered for using virtualization technologies and encrypted SSL/TLS channels via the web administration console.

D. Cryptographic Techniques in Cloud Computing

Challenges to data exposure are intercepted with the use of convergent encryption, homomorphic encryption, and proxy-re-encryption. Identity-Based Cryptography (IBC) and Attribute-Based Cryptography (ABC) serve as potential cryptographic solutions for unauthorized access [3]. However, data loss and manipulation are mitigated using Proof of Data Possession (PDP) and Proof of Retrievability (POR) while Privacy breaches' cryptographic solutions are searchable

encryption and are achieved using Private Information Retrieval (PIR). PDP checks if remote cloud servers have outsourced data using a challenge-response protocol. A client using PDP can check if a file is stored and is available on a cloud server in its original form using a four-step procedure: pre-process, challenge, proof, and verification. POR verifies data integrity and data recoverability in case of failure by adding sentinels in data. The data owner can send a challenge to the cloud server using randomly selected sentinel positions in the data and as a response, the corresponding sentinels must be sent back. If it is not the case, the data are suspected to be modified or deleted.

Searchable Encryption allows a cloud server to search encrypted data using information that the data owner or client has previously provided. This is called a trapdoor [3]. The cloud host hence does not know the exact query nor the data that matches the query thus providing confidentiality with searching ability. The data owner or a cloud client can receive the data locally and decrypt it while saving bandwidth. There are two types of searchable encryptions: symmetric and public keys. PIR schemes provide clients with the ability to request data from cloud servers without revealing which item is being retrieved to the storage itself hence protecting curious cloud providers. PIR is available in two schemes: computation PIR (cPIR) which provides privacy from computationally linked servers and information theoretic PIR (itPIR) which provides privacy for computationally independent servers.

Convergent encryption, a content hash keying cryptosystem, both protects data outsourced to the cloud and ensures client-side data duplication so that the storage can host only one copy of a file regardless of the number of users accessing it. In client-side deduplication, convergent encryption provides two levels of encryption, symmetric data encryption level, and asymmetric key encryption level. At the data encryption level, an enciphering key, K , is derived from the data itself using a one-way hash function and used to encrypt the data so that the same data encrypted by several users will produce the same encrypted data, which will be stored only once [3]. At the asymmetric key encryption level, the depositor uses the recipient's public key and asymmetric encryption to encrypt the key K , to be shared alongside user metadata.

Homomorphic encryption algorithms allow third parties to perform deterministic computations on encrypted data to ensure privacy preservation [3]. Homomorphic mechanisms also allow private queries whereby the client sends an encrypted query and the cloud server replies with an encrypted response without looking at the query itself. A user can also save encrypted data on the cloud and then have the cloud server retrieve only some files that when decrypted satisfy some conditions without the server having decrypted the data. To provide data secrecy against cloud providers, proxy re-encryption algorithms usage has officiated the cloud storage outsourcing storage clouds use proxy re-encryption algorithms to provide data secrecy against cloud providers [3]. If a cloud entity wants to access cloud data from a depositor, the cloud server must first re-encrypt the data using the cloud entity's public key and the server's public master key while considering privileges granted.

To secure user data storage and access on the cloud, Pavani et al. [15] performed a comprehensive review of several types of Attribute-Based Encryption (ABE) where each user is identified by a set of attributes. For key policing, by assigning a range of attributes to each user within a control tree, Encryption Key Policing attributes (KP-ABE) are used in general. En route towards the use of non-monotonic access structures, that used control doors such as NOT operations in the control framework, the Expressive Main Regulation ABE (EKP-ABE) which is an expanded version of Key-Policy ABE (KP-ABE) has been proposed. To protect the user's privacy, Ciphertext-Policy ABE (CP-ABE), an inverse iteration of KP-ABE uses a series of attributes for the user's private key to feed the ciphertext to an access framework. To integrate multiple data domains and associated processes, Hierarchical Attribute Dependent Encryption (HABE) ensured data consistency and accuracy throughout the system.

To protect user revocation, Multi-Authority ABE (MA-ABE) uses secured encryption that processes encryption using global public parameters. Additionally, for file protection, File Hierarchy ABE (FH-ABE) encryption scheme is used where files are protected based on the layered entry layout. Pertaining to encryption where users encrypt data based on attributes induced by measuring user's characteristics and corresponding assigned weights, Ciphertext-Policy Weighted ABE (CPW-ABE) was suggested as an efficient ABE. Moreover, to prevent the disclosure of data owners' and data users' sensitive data, Policy Hidden ABE (PH-ABE) has been used [16]. Multi-level ABE (ML-ABE) allows data to be encrypted through multi-level access control schemes whereby users can access only parts of data. Partially hidden access policies are maneuvered to hide the private information attributes and fully hidden ABE policies are used to hide private information and their values.

Policy Hidden Outsourced ABE (PHO-ABE) allows a user to delegate the decryption process execution to a semi-trusted server but keeps the ability to verify decrypted data correctness. To perform user authentication, Attribute-Based Signatures (ABS) have been proposed whereby the user must hold a set of attributes satisfying an access policy to sign a message. An attribute authority generates user attributes and private keys, and a verification entity verifies generated signatures [16]. Multiple Authority ABS (MA-ABS) allows multiple authorities to manage attributes and private keys. Attribute-Based SignCryption (ABSC) logically combines ABE and ABS in one-step to provide fine-grained and granular access control, authentication of data origin, and data confidentiality.

There is a certain cryptographic method, the location-based encryption method, which allows encryption and decryption to be possible only at a specific location using location details to generate cryptographic keys. Relatively, a hybrid algorithm, comprising of both the symmetric Advanced Encryption Standard (AES) algorithm to encrypt data and the asymmetric Rivest-Shamir-Adleman (RSA) algorithm to encrypt the AES private key for key exchange are used to protect data [17]. The fast computation of symmetric algorithms and the high security of asymmetric key pairs prevent HTTP-centric brute force attacks while guaranteeing secure key exchange.

Cognitive cryptography is used for intelligent data management which involves managing strategic, confidential, and secret data, following protocols to verify every information holder, managing semantic information that data contains, and managing data at all the operational levels of the organization entity. In cognitive cryptography, data are secured using unique personal information from biometrics and using semantic information that distinctly identifies individual features of a participant [18]. The cognitive cryptographic protocol involves concealing and encrypting data by splitting and distributing it among a selected group of secret and trusted entities that are identified and verified using their biometric characteristic information and their selected personal features' semantic description.

The review indicates that there are significant research gaps given the security impacts and issues on cloud-based e-learning platforms. e-Learning technologies deal with end-user data such as personal information, location, behavioral patterns, digital footprints, and intellectual properties such as applications, notes, videos, and research. Since cloud computing deals with the transmission, storage, processing, and access of this e-learning data, the analysis and recommendations cover a broad view and multiple aspects of the entire cloud-based e-learning service.

III. METHODOLOGY

This section demonstrates the methodology used for reviewing the literature of e-learning and cloud computing security and also on the analysis of the identified concerns. An empirical study was performed using qualitative techniques and systematic review as the research instruments for the technical analysis of documents. A systematic review was employed in selecting and critically appraising research relevant to the domain. The documents focus on e-learning technologies, cloud computing, information security, security impacts, and cryptographic techniques on e-learning and cloud computing. These technical documents have been reviewed to learn about how cloud-based e-learning technologies work and about the latest advancements in security aspects and cryptographic techniques in cloud computing and e-learning. The development of cryptography in the recent years related to the theme of cloud-based e-learning is evaluated from the literature using chronological and thematic methods. The results are categorised thematically based on functionalities, security aspects and different cryptographic techniques. Based on the evaluation, several criterias have been thematised to propose an implementation framework in which a critical appraisal has been performed on these findings in order to formulate recommendations based on different scenarios.

IV. RESULTS AND DISCUSSION

Following the study of the state of art in cryptographic techniques applied to cloud-based e-learning systems, the following observations have been made:

1) The entire end-to-end high-level architecture of the cloud-based e-learning service that includes the end-user, the administrators, the internetwork, and the cloud service including servers and storage must be considered when

analyzing the cyber security threats faced and the cryptographic techniques used to protect the various aspects of information.

2) The aspects that require protection are Confidentiality, Integrity, Availability, Authenticity, Accuracy, and Non-Repudiation.

3) We propose to use a hybrid form of hierarchical and distributed approaches to break down each component of the end-to-end cloud-based e-learning system in a top-down manner and analyze each component separately.

4) The end-to-end cloud-based e-learning system can be broken down into three tiers, namely, User Tier, Network Tier, and Cloud Tier.

5) The user tier consists of the end-user devices and services accessed by a web browser or an e-learning application. End users can be students, teachers, or administrators.

6) The network tier consists of the internetwork services provided by a network provider or an internet service provider. Usually, TLS/SSL and IPSec VPN technologies are used to secure the connection.

7) The Cloud tier embodies the server and storage infrastructure that is broken down based on IaaS, PaaS, and SaaS service models. The cryptographic techniques specifically used in the Storage component of the Cloud Tier are also analyzed as an independent entity aside from service models.

8) Each tier has its functionality and security aspect requirements, which are fulfilled by cryptographic techniques.

9) Several criteria have been extrapolated from information gathered and are considered when using cryptographic techniques to fulfill the requirements.

10) While the use of cryptographic techniques has been theorized and implemented in some capacity, no explicit method or framework has been defined for their implementation across the whole cloud-based e-learning spectrum.

A. Cryptographic Techniques Tabulation

These ten observations exhibit several inferred criteria, and the different cryptographic techniques used in the User Tier and the Cloud Tier as tabularized below. The Cloud Tier includes the Storage, IaaS, PaaS, and SaaS cloud computing services. Table I depicts the functionality of end-user devices and security aspect requirements when accessing a web browser or an e-learning platform that uses cryptographic techniques.

Table II represents the Storage Cloud Tier to analyze the storage component of cloud-based services to protect the various aspects of information.

Table III depicts the IaaS Cloud Tier that explores the security aspects of the running applications and different workloads in the cloud.

Table IV illustrates the PaaS Cloud Tier to analyze the security aspects of developing and managing application functionalities and the corresponding cryptographic techniques.

TABLE I. USER TIER

Functionality	Cryptographic Technique	Security Aspect
1. Access Control and Trust Establishment	<ul style="list-style-type: none"> TrustBAC model TSR Model PKI 	Confidentiality Integrity Availability Authenticity Non-Repudiation
2. Authentication	<ul style="list-style-type: none"> Biometrics driven ABC 	Confidentiality Integrity Authenticity
3. Course Content Security	<ul style="list-style-type: none"> Trust Certificates 	Confidentiality Integrity Authenticity Non-Repudiation
4. E-Assessment	<ul style="list-style-type: none"> Location-Based Cryptography Pseudonymization 	Confidentiality Integrity Availability Authenticity Accuracy
5. Usage Anonymity	<ul style="list-style-type: none"> Pseudonymous credentials (ABC) Location-Based Cryptography 	Confidentiality Accuracy
6. Anonymous Search	<ul style="list-style-type: none"> Searchable Encryption 	Confidentiality Accuracy
7. Anonymous Retrieval	<ul style="list-style-type: none"> PIR 	Confidentiality Availability Accuracy
8. Client-Side Deduplication	<ul style="list-style-type: none"> Convergent Encryption 	Confidentiality Integrity Availability Accuracy
9. Client-Side Availability Check	<ul style="list-style-type: none"> PDP 	Confidentiality Availability Accuracy
10. Client-Side Integrity and Recoverability Check	<ul style="list-style-type: none"> POR 	Confidentiality Integrity Availability

TABLE II. STORAGE TIER

Functionality	Cryptographic Technique	Security Aspect
1. File System Security	<ul style="list-style-type: none"> IPFS hash functions (SHA-1, SHA-256, and BLAKE2) 	Confidentiality Integrity Availability Accuracy
2. Data Deduplication	<ul style="list-style-type: none"> Convergent Encryption 	Confidentiality Integrity Availability
3. Data Privacy Preservation	<ul style="list-style-type: none"> Homomorphic Encryption 	Confidentiality
4. Data Secrecy	<ul style="list-style-type: none"> Proxy Re-encryption 	Confidentiality
5. Data Availability Check	<ul style="list-style-type: none"> PDP 	Confidentiality Availability Accuracy
6. Data Integrity and Availability Check	<ul style="list-style-type: none"> POR 	Confidentiality Integrity Availability Accuracy
7. Anonymous Search	<ul style="list-style-type: none"> Searchable Encryption 	Confidentiality Accuracy
8. Anonymous Retrieval	<ul style="list-style-type: none"> PIR 	Confidentiality Accuracy
9. Data Signing	<ul style="list-style-type: none"> ABSC 	Integrity Authenticity Non-Repudiation

10. Intelligent Data Management	<ul style="list-style-type: none"> Cognitive Cryptography 	Confidentiality Integrity Availability Authenticity Non-Repudiation Accuracy
---------------------------------	--	---

TABLE III. IAAS TIER

Functionality	Cryptographic Technique	Security Aspect
1. Access Control and Trust Establishment	<ul style="list-style-type: none"> NTS-based Cyber Trust 	Confidentiality Availability Authenticity Non-Repudiation Accuracy
2. Virtual Machine and Block Data Deduplication	<ul style="list-style-type: none"> Convergent Encryption 	Confidentiality Availability Authenticity
3. Storage Replication for Disaster Recovery	<ul style="list-style-type: none"> PHO-ABE 	Confidentiality Availability Accuracy
4. Virtual Machine Encryption	<ul style="list-style-type: none"> MA-ABE 	Confidentiality Availability Accuracy

TABLE IV. PAAS TIER

Functionality	Cryptographic Technique	Security Aspect
1. Authentication of different types of users	<ul style="list-style-type: none"> HABE 	Confidentiality Availability Authenticity
2. User-based Encryption	<ul style="list-style-type: none"> CPW-ABE 	Confidentiality Availability Accuracy
3. Data Replication between Cloud Providers for High Availability	<ul style="list-style-type: none"> PHO-ABE 	Availability Accuracy

Table V describes the SaaS Cloud Tier to inspect the security aspects of ready-to-use cloud-hosted application functionalities and the corresponding cryptographic techniques.

TABLE V. SAAS TIER

Functionality	Cryptographic Technique	Security Aspect
1. Authentication, Authorization, and Access Control	<ul style="list-style-type: none"> RBAC KP-ABE CP-ABE 	Confidentiality Availability Authenticity
2. File Storage Encryption	<ul style="list-style-type: none"> FH-ABE 	Confidentiality Accuracy
3. Policy-Based Encryption	<ul style="list-style-type: none"> PH-ABE 	Confidentiality Accuracy
4. Object Storage Encryption	<ul style="list-style-type: none"> ML-ABE 	Confidentiality Accuracy
5. File Level Replication	<ul style="list-style-type: none"> PHO-ABE 	Confidentiality Availability Accuracy
6. File Signing	<ul style="list-style-type: none"> ABSC 	Integrity Authenticity Non-Repudiation Accuracy

B. Criteria Definition

The criteria below define what cryptographic techniques need to provide and what can be used to measure and evaluate the application of cryptography in fulfilling the requirements previously mentioned. The criteria consider the security needs of end-users of cloud-based e-learning technologies while entirely leaving the implementation specifics upon the cloud-based e-learning provider.

1) Authentication Criteria: The authentication mechanism should use cryptographic personalized based on the attributes of the users so that third-party attributes do not define the key generated per user.

2) Access Control Criteria: Access control and trust establishment cryptographic techniques should be used granularly based on user identity and attributes following the principles of least privilege and separation of privileges principle.

3) Pseudonymisation Criteria: Cryptographic techniques should ensure that a user's e-learning activity and data cannot be used for traceability and inference basis.

4) Anonymous Searching Criteria: Cryptographic techniques should ensure that any e-learning user can perform searches on encrypted data stored on the cloud without revealing what the original data is, to protect intellectual property and integrity.

5) Anonymous Retrieval Criteria: Cryptographic techniques should ensure that any e-learning user can retrieve encrypted data and part of encrypted data on the cloud without revealing what the original data is, to protect intellectual property and integrity.

6) Anonymous Storage Criteria: Cryptographic techniques should ensure that any e-learning user can store encrypted data stored on the cloud without revealing what the original data is.

7) Deduplication Criteria: Cryptographic techniques should ensure that data is not duplicated when stored on the cloud, whether at the source client-side or the destination cloud side.

8) Replication Criteria: Cryptographic techniques should ensure the high availability of data stored on the cloud through replication without revealing what the original data is.

C. Proposed Implementation Framework

The application and implementation of the above cryptographic techniques along the end-to-end cloud-based e-learning architecture can be done in a holistic and guided software framework as follows:

1) A multi-agent system consisting of software agents could be installed on all users' devices including computers, smartphones, and tablets to perform various activities based on each user's roles and attributes.

2) Administrators could manage cryptography applied to cloud storage and each of the cloud service models of IaaS, PaaS, and SaaS.

3) Students could encrypt, decrypt, sign, authenticate, manage keys, choose ciphers, verify certificates and signatures and check e-assessment cryptography features based on the roles, privileges, and attributes they have been assigned.

4) Since high-end and computationally intensive cryptographic techniques have been proposed, the underlying cloud infrastructure could be provisioned with adequate resources such as Graphical Processing Units (GPU), virtual GPUs (vGPU), and memory, to handle both Cloud Tier cryptographic processes and to sustain client-side ones that could be outsourced to them via the agents.

5) The locally installed software agents could use the available end-user device resources to perform calculations if the device can support such activities using inbuilt technology such as trusted platform modules in the case of computers and laptops or encryption chips for smartphones.

6) The local agents could interface with web browsers installed on the end-user devices and provide access to the e-learning services through a web portal without affecting user experience or adding ambiguity.

7) If the e-learning platform being accessed provides a VPN to the end-user to access its services, the software agent could be the local endpoint of the VPN to the user, providing both token generation and authentication interface and web access to the service via web browser integration.

8) The software agent could also automatically or interactively install security certificates in the web browsers to which it is integrated.

9) The software agent could be integrated into the cloud service management console provided to the administrators by cloud providers via their application programming interfaces.

10) Finally, the administrators could implement and administer the PKI that governs the certificates being used using the software agent.

V. CONCLUSION

The use of cloud computing and its service models to deliver e-learning technologies, followed by the impact on cyber security aspects and threats was reviewed systematically. The innovative cryptographic techniques used in cloud computing and e-learning have been presented in terms of authentication, access control, pseudonymization, data storage, and access. It was observed that instead of a holistic approach to cryptography, only specific aspects of cloud-based e-learning were focused on and that no explicit implementation guidelines nor framework currently exist. As a result, this comprehensive review provides a holistic tabulation of cryptographic techniques for cloud-based e-learning. It also defines a set of criteria that can be used to evaluate whether the existing cryptographic techniques are fulfilling the requirements as needed. Finally, a framework is proposed to implement cryptographic techniques in a unified way across an end-to-end cloud-based e-learning architecture using multi-agent software. The empirical results are considered in the light of some limitations to the existing literature in cryptography

for cloud-based e-learning technologies. For future work, a theoretical framework will be implemented via a holistic approach using different cryptographic techniques. This will lead the pathway to practical implementations of the framework.

REFERENCES

- [1] A. Chopra and A. Chopra, "Security Threats and Remedies in E-Learning System," *International Journal of Computer Science and Telecommunications*, vol. 7, no. 7, pp. 6-10, 2016.
- [2] M. S. Malhi, U. Iqbal, M. M. Nabi, and M. A. Malhi, "E-Learning Based on Cloud Computing for Educational Institution: Security Issues and Solutions," *International Journal of Electronics and Information Engineering*, vol. 12, no. 4, pp. 162-169, 2020.
- [3] N. Kaaniche and M. Laurent, "Data security and privacy preservation in cloud storage environments based on cryptographic mechanisms," *Computer Communications*, vol. 111, pp. 120-141, 2017.
- [4] M. Bosamia and A. Patel, "An overview of cloud computing for e-learning with its key benefits," *International Journal of Information Sciences and Techniques*, vol. 6, pp. 1-10, 2016.
- [5] A. Rahman, S. Sarfraz, U. Shoaib, G. Abbas, and M. A. Sattar, "Cloud based E-Learning, Security Threats and Security Measures," *Indian Journal of Science and Technology*, vol. 9, no. 48, pp. 1-8, 2016.
- [6] M. Bhatia and J. K. Maitra, "E-Learning Platforms Security Issues and Vulnerability Analysis," In: *International Conference on Computational and Characterization Techniques in Engineering & Sciences (CCTES)*, Lucknow, India, pp. 276-285, 2018.
- [7] C. Kiennert, P. Rocher, M. Ivanova, A. Rozeva, M. Durcheva, and J. Garcia-Alfaro, "Security Challenges in e-Assessment and Technical Solutions," In: *21st International Conference on Information Visualisation*, London, UK, pp. 366-371, 2017.
- [8] S. Ramjan, "E-Learning Security for Collaborative Academy in Area of ASEAN Community," In: *12th International Conference on eLearning for Knowledge-Based Society*, Thailand, pp. 23.1-23.8, 2015.
- [9] O. Adejo, I. Ewuzie, A. Usoro, and T. Connolly, "E-Learning to m-Learning: Framework for Data Protection and Security in Cloud Infrastructure," *International Journal of Information Technology and Computer Science*, vol. 4, pp. 1-9, 2018.
- [10] M. Durairaj and A. Manimaran, "A Study on Security Issues in Cloud based E-Learning," *Indian Journal of Science and Technology*, vol. 8, no. 8, pp. 757-765, 2015.
- [11] L. Q. Huan, D. Nyugen, H. Pham, and N. Huynh-Tuong, "Authentication in E-Learning Systems: Challenges and Solutions," *Science and Technology Development Journal - Engineering and Technology*, vol. 3, no. 1, pp. 95-101, 2020.
- [12] A. Kassid and N. El Kamoun, "Towards a new access control model based on Trust-level for E-learning platform," *Journal of Information Assurance and Security*, vol. 11, no. 6, pp. 302-310, 2016.
- [13] D. Melnikov, V. Petrov, N. Miloslavskaya, A. Durakovskiy, and T. Kondratyeva, "Cybertrust in e-Learning Environment based on Network Time Synchronization," In: *8th International Conference on Computer Supported Education*, Setubal, Portugal, pp. 402-407, 2016.
- [14] R. Ali and H. Zafar, "A Security and Privacy Framework for e-Learning," *International Journal for e-Learning Security*, vol. 7, no. 2, pp. 556-566, 2017.
- [15] V. Pavani, P. S. Krishna, A. P. Gopi, and V. L. Narayana, "Secure data storage and accessing in cloud computing using enhanced group-based cryptography mechanism," in: *Materials Today: Proceedings*, 2020.
- [16] S. Belguith, N. Kaaniche, and M. Hammoudeh, "Analysis of Attribute-Based Cryptographic Techniques and their Application to Protect Cloud Services," *Transactions on Emerging Telecommunications Technologies*, e3667, pp. 1-13, 2019.
- [17] N. S. M. Shamsuddin and S. A. Pitchay, "Location-Based Cryptographic Techniques for Data Protection," *Malaysian Journal of Science, Health & Technology*, vol. 4, pp. 65-68, 2019.
- [18] M. Ogiela and L. Ogiela, "Cognitive cryptography techniques for intelligent information management," *International Journal of Information Management: The Journal for Information Professionals*, vol. 40(C), pp. 21-27, 2018.

Various Antenna Structures Performance Analysis based Fuzzy Logic Functions

Antenna Performance Analysis Based FLF

Chafaa Hamrouni^{1*}, Aarif Alutaybi²

Department of Computer Sciences
Taif University-Khurma University College, Zip Code:2935
Khurma, Kingdom of Saudi Arabia

Slim Chaoui³

Department of Computer Engineering and Networks
College of Computer and Information Sciences
Jouf, Kingdom of Saudi Arabia

Abstract—The antenna is a critical component of the communication system. The antenna is used in wireless communication for signal transmission and reception over long distances. There are numerous sorts of antennas, such as wire antennas, traveling wave antennas, reflector antennas, microstrip antennas, and so on. The application of antennas is determined by the antenna's attributes as well as the frequency range of operation. As a result, it is vital to understand the behavior of antennas over a wide range of operations and select the optimum antenna for the application. The performance parameters of the antenna determines its efficiency. VSWR, Return Loss, Directivity, Bandwidth, and more parameters are available. As a result, one of the primary areas of focus is antenna analysis. In this study, we simulate various antenna types and derive performance parameters such as return loss, directivity, and so on. MATLAB will be used to simulate the antenna at various frequencies. When all of the parameters are taken into account, the analysis becomes quite tough. In this case of ambiguity, we use fuzzy logic to calculate the antenna's performance index. A variety of antenna parameters will be fed into the fuzzy inference system, which will make a judgment based on a set of rules. The crisp numbers are turned into fuzzy values using the fuzzification process, then evaluated and defuzzified to obtain the antenna's performance index. The fuzzy inference system will be developed in MATLAB, and the overall system will be modeled in Simulink.

Keywords—Antenna; antenna element; function; fuzzy logic function; fuzzy inference system; Matlab; Simulink

I. INTRODUCTION

A radiated element is a type of electrical equipment that converts electric power into radio waves, allowing the signal to be transmitted over open space. It also converts incoming radio signals to electrical impulses. As a result, in the field of wireless communication systems, the antenna is extremely significant. Several things influence antenna selection. The frequency of operation [1], as well as the application, are two of these criteria. The antenna's performance is determined by numerous criteria such as return loss, reflection coefficient, and voltage standing wave ratio. In this research, multiple antenna topologies are modeled for performance characteristics for different frequencies using MATLAB's antenna [2] toolbox. These parameters are fed into the fuzzy inference system, which analyzes the antenna's performance while taking all of the parameters into account. The fuzzy inference system is created using a rule set drawn from antenna experts'

knowledge. Using fuzzy rules, the crisp values are fuzzified [3] and assessed for performance, while the linguistic values are defuzzified to crisp values. The proposed fuzzy inference method is used to analyze the antenna's performance. Simulink is used to model the system, and the performance of the antennas at various frequencies is assessed, as well as the performance variation of the antennas regarding frequency, is plotted. This study addresses the construction of a system that uses fuzzy logic to examine the performance of antenna configurations. The antenna structures are simulated in this step using the MATLAB antenna design toolkit, and the performance parameters are extracted. These collected parameters are fed into the analyzer, which determines the antenna's performance. Lotfi A. Zadeh published a study on fuzzy logic in 1964. Zadeh continued to develop the fuzzy set theory between 1965 and 1975. Fuzzy logic arose as a result of the challenges experienced by standard mathematical techniques in constructing and evaluating complicated systems. The performance study of the antenna structure is particularly complex because the parameters that determine performance must be considered and studied at the same time. The fuzzy inference system described in this paper was designed to avoid this complexity [4]. The first research team considers the use of fuzzy logic in educational institutions. It discusses how fuzzy logic can be used to analyze student performance. The marks earned by pupils are fed into the fuzzy inference system, which evaluates the student's performance. The rectangular microstrip antenna [5] was modeled and simulated. It also provided a method for integrating Matlab into Visual Basic. Matlab is a great tool for designing and simulating antenna structures of all types. The tool provides a detailed explanation of the performance parameters of the designed antenna. [6] Proposes utilizing Matlab to construct and analyze a parabolic reflector. [7] examines the analysis of E-plane and H-plane normalized patterns. Analysis of the parabolic reflector, such as f/D, gain, and radiation patterns, was performed, and the appropriate results were provided. A second study team considers faculty performance evaluation in educational institutions. In this research, we propose the creation of a fuzzy inference system to evaluate the antenna's performance. MATLAB is used to simulate the many types of antennas [8] describes the modeling of a rectangular microstrip antenna using Matlab and Visual Basic. [9]. Shows a Matlab simulation of an NxN antenna array, with the antenna array factor indicated for each person. It demonstrates the effect of increasing the directivity of the

*Corresponding Author - E-mail: cmhamrouni@tu.edu.sa
DOI: 10.14569/IJACSA.2022.0130109

beam array factor of the antenna array with an increase in antenna elements, as well as analyzing the effect of increasing the antenna elements on the array factor.[10]-[11]-[12]-[13] cover the design and improvement of antenna element performance. The design of many antennas is known in the literature. [14] describes the design of a square patch antenna. A stacked square patch slotted broadband microstrip antenna is described in another paper [15].

II. FUZZY LOGIC

Nowadays, fuzzy control is considered an important tool for control, in addition, it is used for helping developers to solve several problems such as designing switched dynamic output for continuous-time. A new type of dynamic output feedback controllers, namely, switched dynamic parallel distributed compensation controllers, is proposed, which are switched by basing on the values of membership functions. For guaranteeing stabilities, and maintaining parameters values, we propose, for the various antenna structures, abased fuzzy logic functions solution to be used for performance analysis. In practice, type-2 fuzzy logic was initially introduced by Zadeh [16]. The presented technique has been proved to be very interesting especially in complex problems which are treating real-world noisy applications [17]. We know that during system development steps, Type-2 Fuzzy Logic (T2FL) defines the same lexicon of the classical type-1 FL as membership functions, rules, norms operations, fuzzification, inference, and defuzzification [18], but those terms have unlike definitions to picture them. The big differentiation between Type-1 and Type-2 FL consists essentially in kind of fuzzy sets and in the output processor step which precedes the defuzzification bloc; the type-1 MFs are certain and crisp, whereas these type-2 are themselves fuzzy; they are represented by a bounded region limited by two MFs, were corresponding to each primary MF (which is in $[0, 1]$), a secondary MF is used to the primary one. With regard to the output processor, in type1 FLSs it is represented just by the known defuzzification process (center of sets...), however, in type-2 FLSs it consists of two components: Type reduction and defuzzification; type reduction makes a reduction from a type-2 fuzzy output sets to type-1 sets and then these reduced sets will be defuzzified to obtain the final crisp outputs. Zadeh pioneered fuzzy logic in the 1960s and 1970s. Fuzzy logic incorporates human knowledge with operational algorithms. The computer may be programmed to work in the same way that the human mind does. Traditional logic and set theory are all about whether something is true or false, white or black, zero or one. Fuzzy logic, on the other hand, accepts all conceivable values.

A. Fuzzy Sets

The fuzzy set notion is simply an extension of the classical set concept. When compared to the classical set, the fuzzy set is substantially larger. The classical set has only a few membership options, such as true or false, '0' or '1'.

B. Fuzzification and Defuzzification

The values must be linguistic in order to be applied to the fuzzy inference system. The degree of membership in the fuzzy set is used to represent these linguistic values. Fuzzification refers to the process of transforming these crisp linguistic values into fuzzy linguistic values. The technique of producing

quantitative outcomes is known as defuzzification. The fuzzy inference system will generate a fuzzy result that will be represented in terms of the degree of membership of fuzzy sets. Defuzzification assigns explicit real values to the membership degrees of fuzzy sets.

III. IMPLEMENTATION

In this article, several antenna topologies are simulated for different frequency ranges using the MATLAB antenna processing tool. With this collection of characteristics, analyzing the performance of the antenna becomes a tiresome task. At this point of uncertainty, the fuzzy logic idea [19] is used to examine the performance of the antennas while taking into account all of the characteristics. The antennas [20] are simulated for frequency ranges ranging from 1MHz to 10MHz, and the performance characteristics are assessed with a fuzzy inference algorithm to provide a performance index. The derived performance index is displayed versus frequency

IV. SIMULATION RESULTS AND DISCUSSIONS

A. Antenna Design and Simulation

Antennas are constructed and simulated in Matlab using the antenna design toolbox. We primarily built and simulated five antenna structures: a bow-tie antenna, a monopole antenna, a dipole antenna, an inverted f antenna, and a helix antenna. The simulation yields parameters such as directivity, VSWR, and reflection coefficients, which determine the antenna's performance. The concentration of radiation in a specific direction is measured by directivity. It specifies the antenna's directionality. Efficiency affects both directivity and gain. Patterns can be used to simply determine directivity. The ratio of maximal radiation intensity to average radiation intensity is defined as directivity. The return loss is another key aspect that influences performance. It is a parameter that reflects how much power is lost. As a result, it is a critical element in determining antenna performance. The simulation's VSWR and reflection coefficients are retrieved and used for further processing.

B. Development of Fuzzy Inference System

- The If and Then set of rules is used to create a fuzzy inference system. The regulations are determined based on professional guidance, taking into account all factors of antenna performance.
- The fuzzy system's output is the linguistic value, which must be translated back to crisp value. Defuzzification is the name given to this type of conversion. Defuzzification strategies include the max membership concept, the centroid method, the weighted average method, the mean max method, the center of sum, the center of the biggest area, and the first (or last) of maxima. The centroid approach is utilized for defuzzification in this article. The centroid approach is also known as the center of gravity method.
- The fuzzy logic toolbox is used to create the fuzzy decision system. Performance characteristics such as VSWR, reflection coefficient, return loss, and directivity is fed into the system. Fuzzification is a

process that converts crisp input values to fuzzy language variables.

- The core of the membership function for the given fuzzy set A is defined as that region of the universe characterized by complete and full membership in A. This means that the core consists of those universe elements x such that $A(x) = 1$. The set of fuzzy linguistic variables is referred to as the fuzzy set A. Triangular membership functions are studied in this study.

C. Parameters of Performance

During the simulation step performance parameters of the antenna are determined and the performance variation value parameters are presented in Table (I) to the Table (V), for bow-tie antenna, a dipole antenna Table (II), inverted f antenna Table (III), a monopole antenna Table (IV), and helix antenna respectively. We simulated the antennas in the frequencies range from 1 Mhz to 10 Mhz.

We presented values in Table (I). In practice, the obtained performance index parameters are optimized due to the fuzzy inference system, we analyze d using the fuzzy rules and the performance index is obtained. Table VI gives the variation of the performance index of various antennas at different frequencies.

These performance parameters are input to the fuzzy inference system and analyzed efficiently by using the fuzzy rules. We storge, at a different frequency, performance index variation of various antennas in the Table (VI). We present a plot of performance index variation in Fig. 7. We need a reasonable separation range than a semantic obfuscation technique. At level 3, semantic obfuscation technique separation range came to 40.6 km, which is a high accomplishment regarding area protection, yet utility of administration is debased, while at that level enhanced semantic obfuscation technique accomplished balance between area protection and administration utility, see Table I.

TABLE I. PERFORMANCE PARAMETERS OF BOW TIE ANTENNA

Reflection Coefficient	0	0	0	-9e-9	-2.2e-8	-4.7e-8	8.8e-8	-1.5e-7	-2.4e-7	-3.6e-7
Return Loss	0	0	3e-9	9.4e-9	2.3e-8	4.8e-8	8.8e-8	1.5e-7	2.4e-7	3.7e-7
Directivity	18.7	-24.7	-28.2	-30.7	-32.7	-34.2	-35.6	-36.7	-37.8	-38.7
VSWR	4.7e11	3e10	5.8e9	1.85e9	7.55e8	3.64e8	1.97e8	1.15e8	7.19e7	4.7e7

TABLE II. PERFORMANCE PARAMETERS OF DIPOLE ANTENNA

Reflection Coefficient	-4.5e-8	-7.2e-7	-3.7e-6	-1.2e-5	-2.8e-5	-5.9e-5	-1.1e-4	-1.8e-4	-3e-4	-4.6e-4
Return Loss	4.5e-8	7.2e-7	3.7e-6	1.2e-5	2.8e-5	5.9e-5	1.1e-4	1.8e-4	3e-4	4.6e-4
Directivity	-18.2	-22.9	-24.9	-26	-26.5	-26.9	-27.2	-31.4	-31.7	-31.9
VSWR	3.6e8	2.4e7	4.75e6	1.5e6	9.1e4	2.9e5	1.5e5	9.1e4	5.7e4	3.7e4

TABLE III. PERFORMANCE PARAMETERS OF INVERTED F ANTENNA

Reflection Coefficient	-1.9e-15	-6e-14	5e-13	1.7e-12	3.9e-12	8.34e-12	-1.5e-11	-2.6e-11	-4.2e-11	-6.3e-11
Return Loss	1.9e-15	6e-14	5e-13	1.7e-12	3.9e-12	8.34e-12	1.5e-11	2.6e-11	4.2e-11	6.3e-11
Directivity	1.74	1.73	1.73	1.73	1.73	1.73	1.73	1.73	1.73	1.73
VSWR	9e15	2.9e14	3.5e13	1.01e13	4.45e12	2e12	1.1e12	6.6e11	4.1e11	2.7e11

TABLE IV. PERFORMANCE PARAMETERS OF MONOPOLE ANTENNA

Reflection Coefficient	-1.37e-8	-2.2e-7	-1.1e-6	-3.5e-6	-8.6e-6	-1.8e-5	-3.36e-5	-5.8e-5	-9.3e-5	-1.4e-4
Return Loss	1.37e-8	2.2e-7	1.1e-6	3.5e-6	8.6e-6	1.8e-5	3.36e-5	5.8e-5	9.3e-5	1.4e-4
Directivity	-18.7	-24.7	-28.1	-30.6	-32.5	-34	-35.2	-36.3	-37.2	-38
VSWR	1.26e9	7.88e7	1.53e7	4.9e6	2e6	9.6e5	5.1e5	3e5	1.8e5	1.2e5

TABLE V. PERFORMANCE PARAMETERS OF HELIX ANTENNA

Reflection Coefficient	-1e-11	-1.6e-10	-8.5e-10	-2.6e-9	-6.5e-9	-1.4e-8	-2.5e-8	-4.3e-8	-6.9e-8	-1.05e-7
Return Loss	1e-11	1.6e-10	8.5e-10	2.6e-9	6.5e-9	1.4e-8	2.5e-8	4.3e-8	1.05e-7	1.05e-7
Directivity	-18.4	-23.8	-26.4	-27.8	-28.7	-29.3	-29.7	-30	-30.2	-30.3
VSWR	1.6e12	1e11	2e10	3.4e9	2.6e9	1.3e9	6.8e8	4e8	2.5e8	1.6e8

TABLE VI. PERFORMANCE INDEXC OF VARIOUS ANTENNA AT DIFFERENT FREQUENCY

Bow Tie	0.43032	0.430105	0.49982	0.4499817	0.4998660	0.499674	0.499878	0.49988	0.49988	0.49988
Invertedf	0.500385	0.585188	0.585288	0.58529	0.585296	0.585295	0.585295	0.585295	0.585295 0	0.585295
Helix	0.500856	0.4998837	0.499659	0.4996323	0.499804	0.4997788	0.49977	0.499792	0.499802	0.499804
Monopole	0.500042	0.499877	0.499824	0.499814	0.499857	0.49987	0.499878	0.49988	0.49981	0.49969

Fig. 1 depicts the membership function for the input reflection coefficient. The fuzzy set includes the variables more negative(mn), negative(n), and zero(z), as indicated in the picture. Fig. 2 and 3 depict the rule viewer for the given set of inputs. Fig. 4 depicts a surface view of the variation of the performance index with regard to the input parameters.

Fuzzy Surface Viewer Showing the Variation of Performance Index with Respect to Inputs:

The generated fuzzy system is exported to Simulink, and fuzzy modeling is performed, as shown in Fig. 5. Fig. 6 depicts the suggested system's modeling.

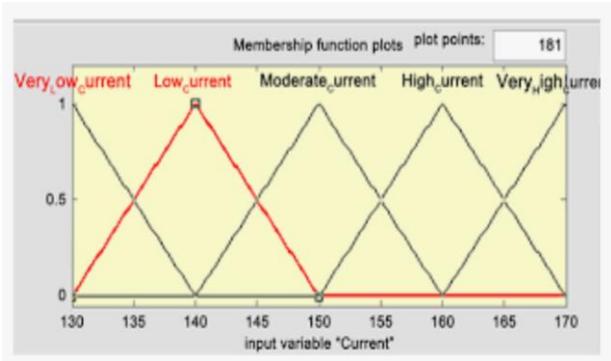


Fig. 1. Membership Function of the Input Variable.

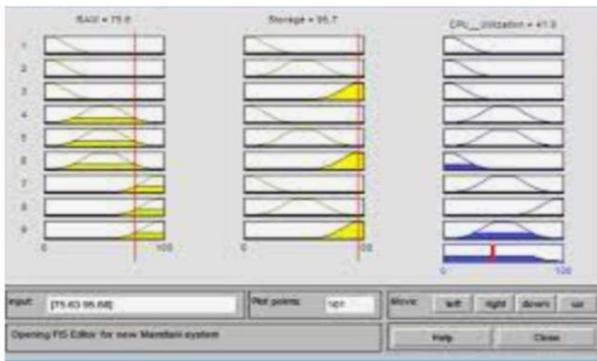


Fig. 2. Fuzzy Rule Viewer for Input VSWr and Directivity of the Input Variable.

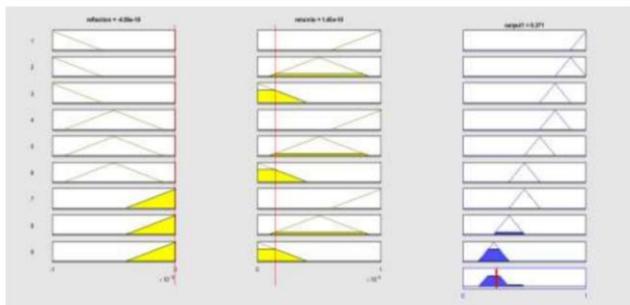


Fig. 3. Fuzzy Rule Viewer for Inputs Reflection Coefficient and Return Loss.

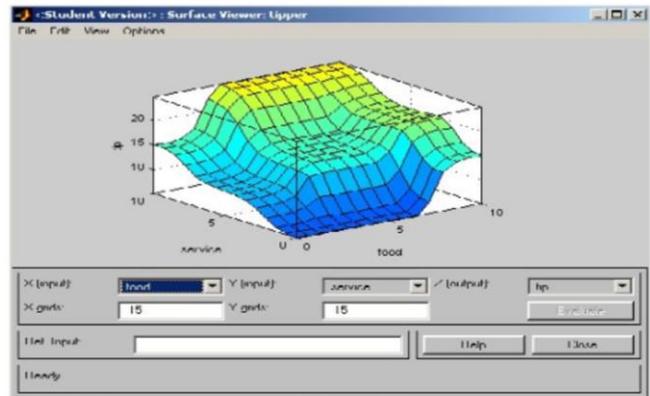


Fig. 4. Fuzzy Surface Viewer showing the Variation of Performance Index with respect to Inputs.

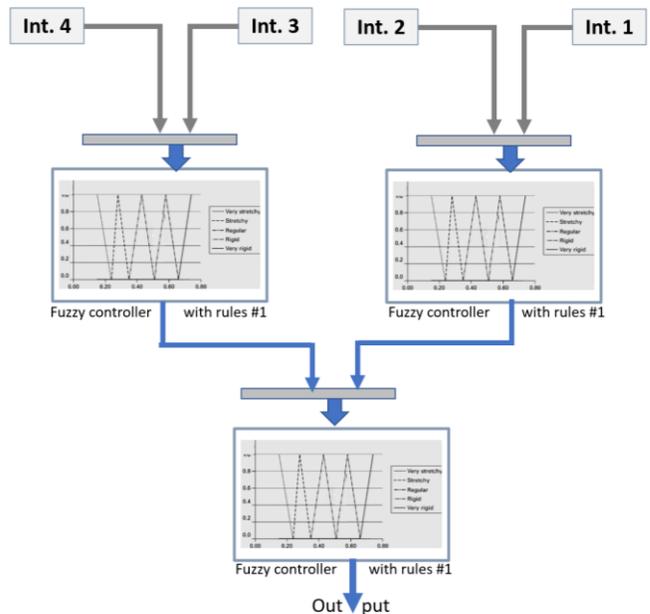


Fig. 5. Simulink Model for Fuzzy.

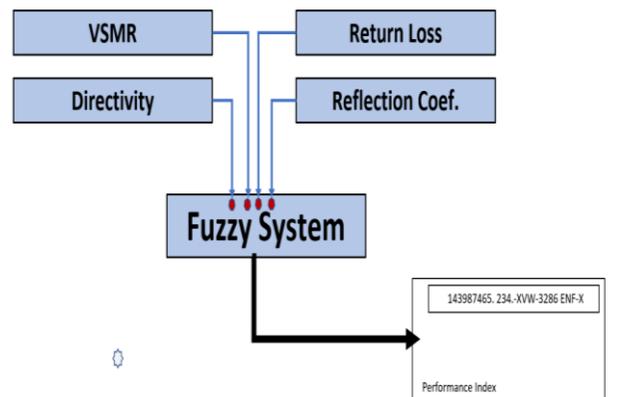


Fig. 6. Proposed System Modeling based on Simulink.

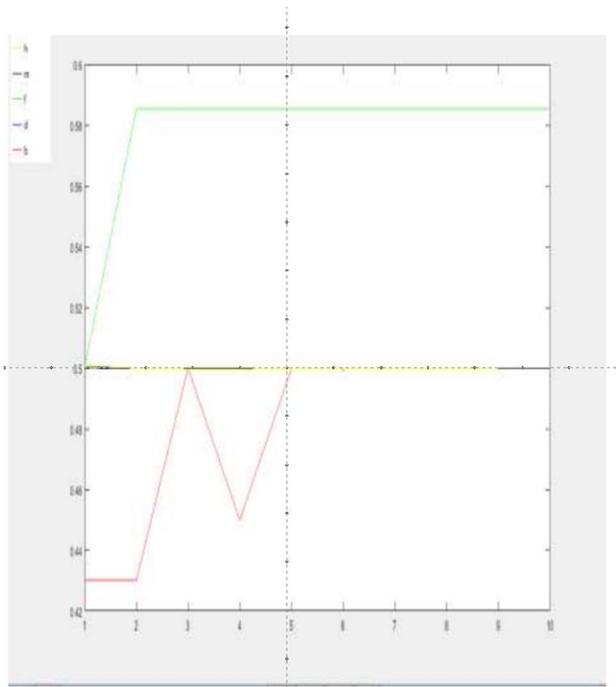


Fig. 7. Performance Index Variation with to Frequency.

Variation is evident when the performance index is plotted against the frequency. Fig. 7 depicts the frequency fluctuation of the performance index.

V. CONCLUSION

In this research, various antenna topologies are simulated using Matlab, and performance characteristics are retrieved from the simulation results. The antennas are simulated for frequency ranges ranging from 1MHz to 10MHz. Based on fuzzy logic, this research provides a new method for evaluating the performance of an antenna. The traditional mathematical paradigm for decision making cannot be applied to complicated systems such as antenna performance. As a result of this paper, this problem is alleviated, and an exact evaluation of the antenna's performance is possible. The fuzzy system is created, and the proposed system is modeled with Simulink. The constructed system is fed inputs, and the output is observed. The obtained performance index of the antennas is plotted. According to the results of the analysis, the inverted f antenna outperforms the other antennas studied in the frequency range 1MHz to 10MHz.

ACKNOWLEDGMENT

The Author would like to acknowledge the Dean of the Khurma University College and the Taif University Department of Scientific Research in the Kingdom of Saudi Arabia, for motivation to accomplish the research work.

REFERENCES

[1] Q. Li, Y. Tian, Y. Zhang, L. Shen and J. Guo, "Efficient Privacy-Preserving Access Control of Mobile Multimedia Data in Cloud Computing," *IEEE Access*, vol.7, no.3, pp.131534–131542, 2015.
[2] X. Li, Q. Wang, X. Lan, X. Chen, N. Zhang and D. Chen, "Enhancing Cloud Based IoT Security Through Trustworthy Cloud Service," *IEEE Access*, vol.7, no.5, pp. 9368 - 9383, 2019.

[3] S. Siboni, V. Sachidananda, Y. Meidan, M. Bohadana, Y. Mathov et al., "Security Testbed for Internet-of-Things Devices," *IEEE Transactions on Reliability*, vol. 68, no.1, pp. 23 – 44, 2009.
[4] J. Tang, R. Li, K. Wang, X. Gu and Z. Xu, "A novel hybrid method to analyze security vulnerabilities in Android applications," *Tsinghua Science and Technology*, vol. 25, no.5, pp. 589 – 603, 2020.
[5] P. Li, C. Xu, H. Xu, L. Dong and R. Wang, "Research on data privacy protection algorithm with homomorphism mechanism based on redundant slice technology in wireless sensor networks," *vol.16, no.1*, pp.158 – 170, 2019.
[6] H. Ma, C. Jia, S. Li, W. Zheng and D. Wu, "Dynamic Software Watermarking Using Collatz Conjecture," *IEEE Transactions on Information Forensics and Security*, vol. 14, no.11, pp. 2859 – 2874, 2019.
[7] K. Swamy, S. Wang, T. Bauer, D. Agrawal, A. Abbadi et al. , "Preserving Location Privacy in Geosocial Applications," *IEEE Transactions on Mobile Computing*, vol.13, no.3, pp. 159 – 173, 2012.
[8] M. Yang, T. Zhu, B. Liu, Y. Xiang and W. Zhou, "Differential Private Queries via Johnson-Lindenstrauss Transform," *IEEE Access*, vol. 6, no.5, pp. 29685 – 29699, 2018.
[9] H. Wang, C. Gao, Y. Li, Z.L. Zhang and D. Jin, "Revealing Physical World Privacy Leakage by Cyberspace Cookie Logs," *IEEE Transactions on Network and Service Management*, vol. 17, no.4, pp. 2550 – 2566, 2020.
[10] K.W. Hipel, L. Fang, K. Yang, Y. Chen, "An Interactive Portfolio Decision Analysis Approach for System-of-Systems Architecting Using the Graph Model for Conflict Resolution," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no.10, pp. 1328–1346, 2014.
[11] M.A. Cardin, J. Yixin, H. Yue and F. Haidong, "Training Design and Management of Flexible Engineering Systems: An Empirical Study Using Simulation Games," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no.9, pp. 1268–1280, 2015.
[12] Y. Wei, H.R. Karimi and W. Ji, "A Novel Memory Filtering Design for Semi-Markovian Jump Time-Delay Systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 48, no.12, pp. 2229–2241, 2017.
[13] K. Xing, M.C. Zhou, F. Wang, H. Liu and F. Tian, "Resource-Transition Circuits and Siphons for Deadlock Control of Automated Manufacturing Systems," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 41, no.1, pp. 74 – 84, 2010.
[14] C. hamrouni, "Complex ESP Systems Proposal based on Pump Syringe and Electronically injector Modules for Medical Application," *Journal of Multimedia Information System (JMIS)*, vol.7, no.2, pp.175-188, 2020.
[15] J. Wu, Z. Guang, J. Li, G. Wang; H. Zhao and W. Chen, "Practical Adaptive Fuzzy Control of Nonlinear Pure-Feedback Systems With Quantized Nonlinearity Input," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no.3, pp.638– 648, 2018.
[16] S. Wang, X. Meng, T. Chen, "Orbital Mechanics Wide-Area Control of Power Systems Through Delayed Network Communication," *IEEE Transactions control Systems Technology*, vol. 20, no.2, pp. 495 – 503, 2011.
[17] M. Eisen, M.M. Rashid, K. Gatsis, D. Cavalcanti, N. Himayat et al. , "Control Aware Radio Resource Allocation in Low Wireless Control Systems," *IEEE Internet of Things*, vol. 6, pp. 7878 – 7890, 2019.
[18] C. Lu, A. Saifullah, B. Li, M. Sha, H. Gonzalez et al. , "Real-Time Wireless Sensor-Actuator Networks for Industrial Cyber-Physical Systems," *Proceedings of the IEEE*, vol. 104, no.5, pp. 1013–1024, 2015.
[19] Y.S. Sinem and C. Ergen, "Joint Optimization of Wireless Network Energy Consumption and Control System Performance in Wireless Networked Control Systems," *IEEE Transactions on Wireless Communications*, vol. 16, no.4, pp. 2235– 2248, 2017.
[20] Y. Wang, S.X. Ding, D. Xu and B. Shen, "An Fault Estimation Scheme of Wireless Networked Control Systems for Industrial Real-Time Applications," *IEEE Transactions on Control Systems Technology*, vol. 22, no.6, pp. 2073 – 2086, 2014.

Empirical Analysis Measuring the Performance of Multi-threading in Parallel Merge Sort

Muhyidean Altarawneh¹, Umur Inan²

Department of Computer Science
Maharishi International University, Fairfield, Iowa, USA

Basima Elshqeirat³

Department of Computer Science
University of Jordan, Amman, Jordan

Abstract—Sorting is one of the most frequent concerns in Computer Science, various sorting algorithms were invented for specific requirements. As these requirements and capabilities grow, sequential processing becomes inefficient. Therefore, algorithms are being enhanced to run in parallel to achieve better performance. Performing algorithms in parallel differ depending on the degree of multi-threading. This study determines the optimal number of threads to use in parallel merge sort. Furthermore, it provides a comparative analysis of various degrees of multithreading. The implementation in this empirical experiment takes a group of devices with various specifications. For each device, it takes fixed-sized data set and executes merge sort for sequential and parallel algorithms. For each device, the lowest average runtime is used to measure the efficiency of the experiment. In all experiments, single-threaded is more efficient when the data size is less than 10^5 since it claimed 53% of the lowest runtime than the multithreaded executions. The overall average of the experiments shows either four or eight threads, with 72% and 28%, respectively, are most efficient when data sizes exceed 10^5 .

Keywords—Parallel merge sort; sort; multithread; degree of multithreading

I. INTRODUCTION

Merge sort is a divide and conquer algorithm that was invented by John von Neumann in 1945, it is an efficient, general-purpose, comparison-based sorting algorithm [1]. Most implementations produce a stable sort, which means that the implementation preserves the input order of equal elements in the sorted output. A detailed description and analysis of bottom-up merge sort appeared in a report by Goldstine and Neumann as early as 1948 [2]. Such divide and conquer algorithm recursively break down a problem into sub-problems, making it simple to be solved easily, then combine the solutions of the sub-problems until the original problem is solved. In sorting n objects (list of array elements), merge sort is an efficient algorithm that has an average and worst-case performance of $O(n \log n)$ [2].

If the running time of merge sort for a list of length n is $T(n)$, then the recurrence $T(n) = 2T(n/2) + n$ follows from the definition of the algorithm (apply the algorithm to two lists of half the size of the original list and add the n steps taken to merge the resulting two lists). In the worst case, the number of comparisons merge sort makes is equal to or slightly smaller than $(n \log n - 2 \log n + 1)$, which is between $(n \log n - n + 1)$ and $(n \log n + n + O(\log n))$ [3]. In the section below, a pseudo-

code of merge sort is illustrated, followed by an example in Fig. 1, using a simple data set of $\{38,27,43,3,9,82,10\}$ [4].

Fig. 1 illustrates how the algorithm divides all items one by one then combines them recursively. This approach indicates the possibility of applying the algorithm in parallel. Hence, parallel merge sort reduces the complexity to $O(n \log n / t)$, where t is the number of threads, by using multi-threaded operations where the data is divided into equal portions and each portion is assigned to a specific thread. The complexity is reduced to $O(n)$ but could vary according to the number of threads used [5].

Merge sort is suitable when the data structure is a linked list because it is a sequential access structure. Implementing a linked list hinders the performance of other algorithms such as quicksort and heapsort [6,7]. Moreover, parallel merge sort is frequently used in various domains, including; sorting NoSql databases [8], high-performance computing environments [9], and massively parallel architectures [10,11].

Algorithm 1 Merge Sort

```
1: procedure Mergesort
2:   var list left ,right , result
3:   if length(m) ≤ 1 then return m
4:   else
5:     var middle = length(m) / 2
6:     for each x in m up to middle do
7:       add x to left
8:     end for
9:     for each x in m after middle do
10:      add x to right
11:    end for
12:    left ← mergesort(left)
13:    right ← mergesort(right)
14:    result ← merge(left, right) return result
15:
```

When it comes to executing algorithms in parallel, most studies show results of the performance on several processors [12-15]. These results will mainly rely on the specifications of the device and the behavior of the execution in terms of multithreading. The question that led to this research is, what is the suitable degree of multi-threading required for parallel merge sort? This study conducts an empirical experiment and highlights several factors that influence multithreading performance. First, the number of cores that affect multithreading performance and second, the given data size that demands multithreading when a single-threaded performance degrades.

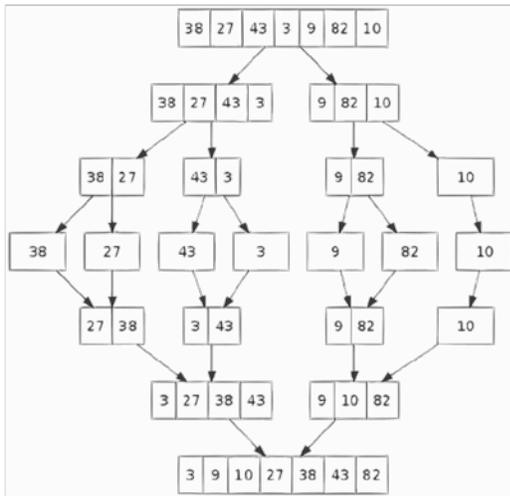


Fig. 1. Merge Sort Algorithm.

The contribution of this paper is to determine the optimal number of threads to use in parallel merge sort. Furthermore, it provides a comparative analysis of various degrees of multithreading. Each data size is examined among a determined number of threads, starting from one thread (sequential), two, four, eight, and sixteen threads (parallel).

In Section 2, related studies were taken to see how parallel merge sort was implemented and what the results were. Section 3 explains and walks through how the experiment was conducted. The results are illustrated in Section 4 and elucidated in the discussion. Finally, Section 5 presents the conclusion of this study.

II. RELATED WORK

There have been several papers that conducted various researches on parallel merge sort, and they have come up with the following.

Jeon [13] improved parallel merge sort by distributing and computing the approximately equal number of keys in all processors throughout the merging phases. Using the histogram information, keys can be divided equally regardless of their distribution, which evaluated the speedup showing a better performance by applying parallel merge sort on two different parallel machines: a Cray T3E and a Pentium III PC cluster on maximum data size of $10^6 \times 4$.

The tested algorithm on loosely coupled parallel machines and the performance of the algorithm has been observed. It has been found that the computational time of the algorithm varies logarithmically for a varying number of processors scenario [14].

Uyar [5] experimented with applying parallel merge sort using multi-threads similar to this experiment. It stated that two threads could perform one merge operation simultaneously. One thread generates the first half of the sorted values that start from the minimums of the two sorted subsets. The other thread generates the second half of the sorted values starting from the maximums of the two sorted subsets. It also compared it with double merging by using four threads implementing it on Java. The comparison focused on array sizes from 10 million up to

50 million. In this study, the array size starts from 5000 up to 50 million to detect when executing in parallel is more efficient than sequential.

A study was conducted on three parallel sorting algorithms (Odd-even transposition sort, Parallel rank sort, and Parallel merge sort) on a number of processors 2, 4, 6, 8, 10, and 12 on 10000 integers [15]. The results proved that parallel merge sort was the fastest, yet the study was comparing only one input size and may differ when the data size increases.

These previous studies show that merge sort could be conducted in parallel in several ways, giving better results than sequential as the array size increases [5,13-15]. Yet, these studies were concerned with enhancing the performance of merge sort without comparing the degree of multi-threading. Only [5] compared different array sizes that were only applied up to four threads on a specific range of sizes, from 10^6 to $10^6 \times 5$. This study experiments parallel merge sort on four different degrees of multi-threading in a broader range of array sizes from 10^5 to 10^7 , which is explained in Section 3 maintaining the integrity of the specifications.

III. EXPERIMENT

A. Requirements

This experiment was implemented on Java SE8. It was conducted on five devices to ensure diversity in the environment of implementation. Moreover, to verify the results are not dependent on the specifications of a particular device. The specifications of the devices used in this experiment are shown in Table I.

B. Implementation

This experiment takes a specific data set and executes it in two approaches: 1) Sequential (one thread), 2) Parallel (two, four, eight, and sixteen threads). The source code is available on <https://github.com/muhyidean/ParallelMergeSort.git>.

The implementation in this experiment takes a data set and applies merge sort for sequential and parallel algorithms. For sequential, it executes Algorithm 1. As for parallel, it executes Algorithm 2 based on the following:

1) *Data formation:* The array sizes for the data sets begin from $10^3 \times 5$, 10^4 , $10^4 \times 5$, 10^5 , ... up to 10^7 . Based on the array size, ten different random data sets are initiated to be implemented in both execution approaches. Each data set will be placed in a separate array and executed in each approach. The average runtime of ten executions for each array size is taken in milliseconds.

TABLE I. DEVICE SPECIFICATIONS

	OS	Processor	# Cores	RAM
Device 1	Windows	Intel i5	4	16
Device 2	Windows	Intel i7	8	16
Device 3	macOS	Intel i5	4	8
Device 4	macOS	Intel i7	8	16
Device 5	Ubuntu	Intel i5	4	4

2) *Partition process*: The partitioning will be in five categories, one in sequential and four degrees of multi-threading 2, 4, 8, 16. The original data set is considered the first partition, so it will be directly executed (sequentially). Then the same data set is taken and split in half making two data sets, each partition is assigned to a thread to run parallel. The process goes on for the other partitions with respect to the number of threads to be implemented which are two, four, eight, and sixteen.

3) *Thread management*: The implementation for the parallel merge sort divides the array into sub-arrays to be sorted by the number of threads. The threads sort their assigned sub-arrays independently. Two consecutive sorted sub-arrays are combined by one thread. Each merging thread merges two sorted arrays. The merge operation follows this approach. Whenever the arrays are sorted, the number of arrays is decreased by half. During the last iteration, two sorted arrays are merged to produce a sorted array. This implementation did not use any third-party libraries/frameworks, it was implemented with the java thread package in JDK (Java Development Kit).

Fig. 2 illustrates the partitioning process and the merging mechanism. Each elliptical shape is considered a thread; the shapes labeled with D represent the partition of the original array sorted by merge sort. The shapes labeled with M merge the results from the previous threads until it merges the whole array. To be better illustrated, sixteen threads are not shown Fig. 2 because it follows similar partitioning.

C. Data Analysis

Tables III to VII shows the average runtime for different array sizes on each. Furthermore, they also show how each device performs on different execution approaches (sequential and parallel). For example, the average execution time is calculated by running the algorithm ten times, then the average of times is taken. Table II is one of the execution results for device 4 on array size 10^5 . For instance, the result shows that (Th-4) was the most efficient for this case. However, it may differ as the size increases and is subject to the device specifications. For each device, on each data size, it will have a table like Table II.

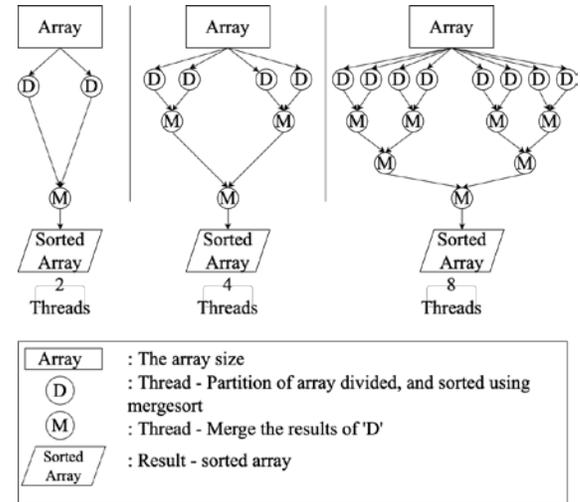


Fig. 2. Parallel Merge Sort using Three Degrees of Multi-threading (2,4,8).

Algorithm 2 Parallel Merge Sort

```

1: procedure PMergesort
2: var val ← (v) // v: the number of values here
3: // x: the number of threads
4: var list arr test _1[ ], arr test 2[ ], ...arr test x[ ] // Defining main arrays
5: var list arr 2[ ] ... arr x[ ] // Defining sub arrays
6: // Defining threads to execute merge sort for each array
7: threads t1(mergesort(arr 1)), t2(mergesort(arr 2))... t_x(mergesort(arr x))
8: // Assign random integers to main arrays, to give each same set of random values
9: for i ← 0 to val do
10: n ← random value in range of (1 - x)
11: arr test1[ ], arr test 2[ ]...arr _test x[ ] ← n
12: end for
13: // Partition data set and add into x sub arrays for each set of threads
14: var mid ← (length of arr test x/x) // Get mid points for each partition
15: * repeat code in line 14 for x partitions
16: // Calculate the time taken for each set of threads
17: var ts ← take current time
18: execute t1 , t2 ... tx // Execute threads
19: var te ← take current time
20: * repeat codes in lines (17 - 19) for each set of threads (2,4, 8 ... x )
21: var tr ← ts - te // to calculate the time taken in parallel mergesort (x threads)
22: file ← export results(tr1,tr2...trx) // to take results (time taken in milliseconds)
23: end procedure=0
    
```

TABLE II. DEVICE 1 – RUNTIME ON SIZE 10^5 (MS)

Execution #	Th-16	Th-8	Th-4	Th-2	Th-1
Execution 1	159	14	18	17	46
Execution 2	33	21	37	50	27
Execution 3	37	22	31	36	36
Execution 4	23	26	30	24	44
Execution 5	32	32	16	32	49
Execution 6	38	48	38	32	32
Execution 7	17	16	32	81	44
Execution 8	31	48	32	33	32
Execution 9	35	33	16	49	34
Execution 10	14	16	25	97	16
Average	41.9	27.6	27.5	45.1	36.0

IV. RESULTS AND DISCUSSION

This section highlights and points out the main findings of the empirical experiment. To measure the efficiency of the experiment, the lowest average execution time (ms) is taken for each data size on each device.

A. Results

In Tables III to VII, it shows the average of 10 executions for each degree of multi-threading. Each column is a different size starting from $10^3 \times 5$ up to 10^7 . The rows show the performance of each thread for a specific data size. For example, (Th-1) is one thread, (Th-2) is two threads and goes on. As shown in Tables III to VII, for data size $10^3 \times 5$, all devices perform efficiently in terms of runtime in a single-threaded execution. As for the sizes 10^4 and $10^4 \times 5$, it varies from one to eight threads depending on the number of cores in the device. With data sizes of 10^5 and larger, each device performs better with a certain number of threads, depending on the number of cores. All results are illustrated in Fig. 3 to 7.

Fig. 3 to 7 illustrates the performance graphs according to different data sizes and the number of threads used. Multithreading is clearly more efficient when the data size increases. The appropriate number of threads will generally be visible when the data size exceeds 10^5 .

B. Findings

There were two main findings from these results. First, multithreading does not always have the most efficient runtime as it depends on the data size. Second, even when the data size increases, a specific number of threads will determine the optimized performance based on the device specifications. In other words, implementing as many threads as possible will not lead to higher runtime performance.

Tables VIII and IX were presented to highlight the findings of the results, one below 10^5 and the other greater 10^5 . Table VIII shows the overall average for each device with data sizes below 10^5 . For example, in Device 1, the sequential runtime performance was most efficient. By taking the overall average,

single-threaded was more efficient since it claimed 53% of the lowest runtime than the multithreaded executions. Table IX shows the overall average for each device with data sizes above 10^5 . As shown in Table IX, multi-threaded implementation with either four or eight threads provided better performance with 72% and 28%. Fig. 8 and 9 visualize which threads performed better in the overall average for different data sizes. A higher percentage indicates that using a specific number of threads is more efficient on a particular data size.

Based on the experiment results, all devices that have four cores achieved efficient runtime performance with four threads. Moreover, all devices with eight cores achieved efficient runtime performance with eight threads. Evidently, the selection of the number of threads is mainly determined by the number of the cores.

C. Discussion

The main question of this study is, what is the optimal number of threads for parallel merge sort considering two main factors: data size and number of cores?

TABLE III. DEVICE 1 - RESULTS - AVERAGE RUNTIME (MS)

Th(x) = number of threads	Array Size							
	$10^3 \times 5$	10^4	$10^4 \times 5$	10^5	$10^5 \times 5$	10^6	$10^6 \times 5$	10^7
Th-16	18	17	25	42	94	165	678	1303
Th-8	10	18	34	28	90	131	588	1173
Th-4	8	13	26	28	83	130	585	1142
Th-2	8	11	38	45	104	179	818	1724
Th-1	2	4	23	36	145	261	1342	2751

TABLE IV. DEVICE 2 - RESULTS - AVERAGE RUNTIME (MS)

Th(x) = number of threads	Array Size							
	$10^3 \times 5$	10^4	$10^4 \times 5$	10^5	$10^5 \times 5$	10^6	$10^6 \times 5$	10^7
Th-16	3	3	5	9	30	51	228	523
Th-8	3	4	8	20	36	44	201	415
Th-4	4	3	7	19	34	47	251	523
Th-2	7	1	9	20	43	69	371	778
Th-1	1	2	9	34	69	134	693	1442

TABLE V. DEVICE 3 - RESULTS - AVERAGE RUNTIME (MS)

Th(x) = number of threads	Array Size							
	$10^3 \times 5$	10^4	$10^4 \times 5$	10^5	$10^5 \times 5$	10^6	$10^6 \times 5$	10^7
Th-16	4	5	10	23	57	97	501	1011
Th-8	11	5	16	13	45	86	431	943
Th-4	4	3	9	10	43	85	420	859
Th-2	2	3	7	13	62	130	665	1356
Th-1	1	3	13	27	114	206	1100	2258

TABLE VI. DEVICE 4 - RESULTS - AVERAGE RUNTIME (MS)

Th(x) = number of threads	Array Size							
	10 ³ x 5	10 ⁴	10 ⁴ x 5	10 ⁵	10 ⁵ x 5	10 ⁶	10 ⁶ x 5	10 ⁷
Th-16	3	3	7	13	27	47	371	671
Th-8	2	3	4	10	23	37	184	474
Th-4	2	2	5	8	32	45	259	604
Th-2	1	2	7	13	44	75	387	801
Th-1	2	2	11	23	93	160	832	1660

TABLE VII. DEVICE 5 - RESULTS - AVERAGE RUNTIME (MS)

Th(x) = number of threads	Array Size							
	10 ³ x 5	10 ⁴	10 ⁴ x 5	10 ⁵	10 ⁵ x 5	10 ⁶	10 ⁶ x 5	10 ⁷
Th-16	6	15	24	29	63	108	446	1198
Th-8	7	15	22	27	88	120	440	1064
Th-4	9	11	21	25	63	104	421	931
Th-2	8	12	20	32	87	120	654	1492
Th-1	2	4	23	36	149	261	1336	2200

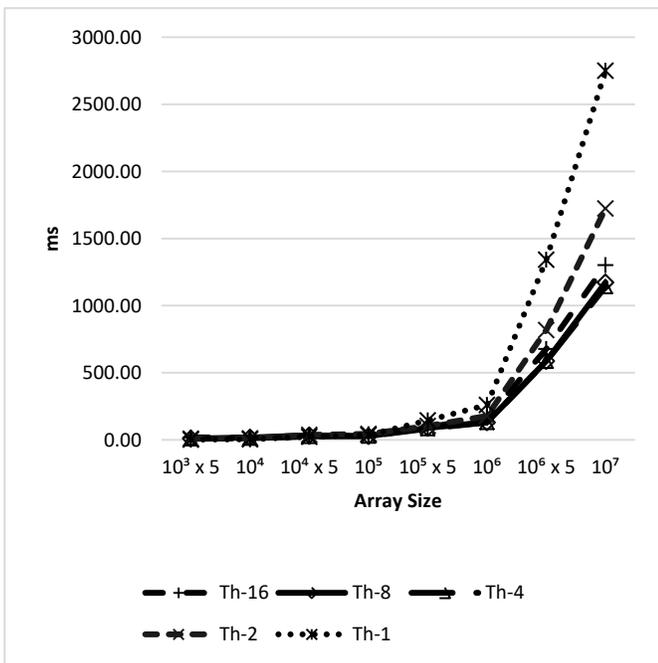


Fig. 3. Device 1 - Results - Average Runtime (MS).

The results of this study had shown that having as many threads as possible will not lead to the best runtime performance. To achieve the best runtime performance, the number of cores present is crucial in determining the optimal number of threads. The cruciality is due to how multiple threads are executed by the operating system. Correspondingly, the data size determines whether multiple threads are required. In small data sets, the use of multiple threads is unnecessary since one thread can perform more efficiently.

The conclusion is that if the data size is under 10⁵, single-threaded will be more efficient. In contrast, having multiple threads will perform better for data sizes that exceed 10⁵. In addition, it should not spawn threads more than the number of cores (excluding merging threads).

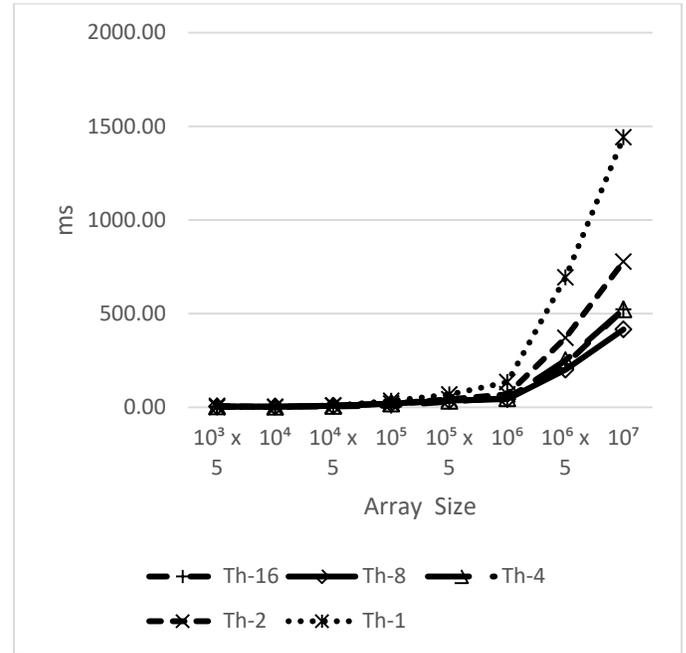


Fig. 4. Device 2 - Results - Average Runtime (MS).

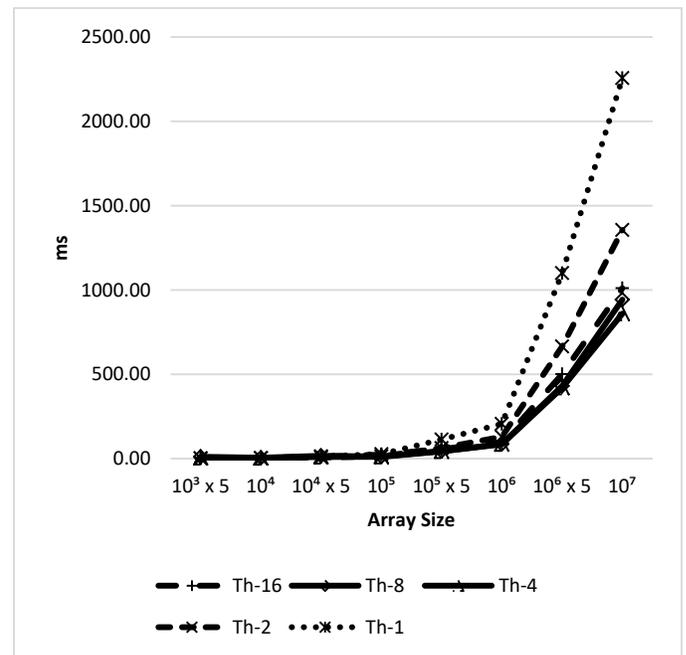


Fig. 5. Device 3 - Results - Average Runtime (MS).

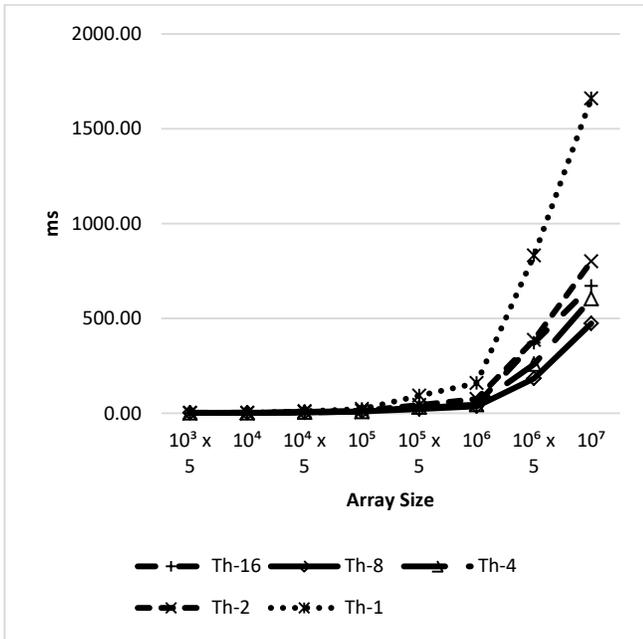


Fig. 6. Device 4 - Results - Average Runtime (MS).

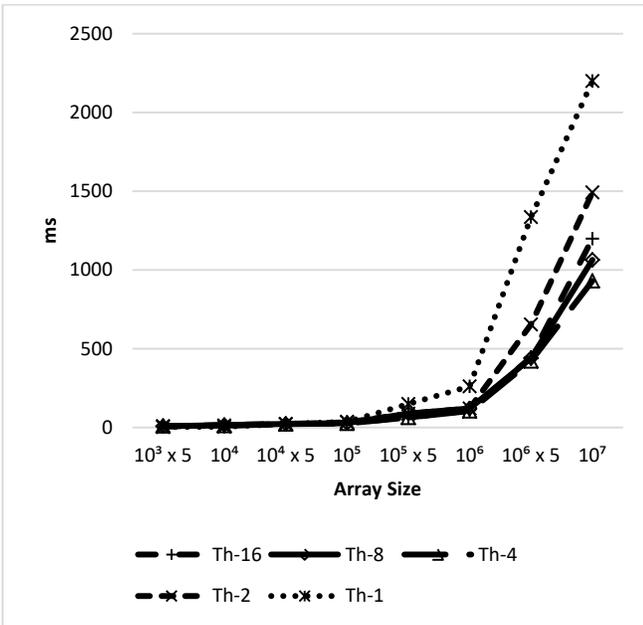


Fig. 7. Device 5 - Results - Average Runtime (MS).

TABLE VIII. MULTITHREADING EFFICIENCY PERCENTAGE (< 50000)

Device	Th(x)= number of threads				
	Th-1	Th-2	Th-4	Th-8	Th-16
Device 1	1.00	0	0	0	0
Device 2	0.33	0.33	0.33	0	0
Device 3	0.33	0.66	0	0	0
Device 4	0	0.66	0	0.33	0
Device 5	1.00	0	0	0	0
Average	0.53	0.33	0.06	0.06	0

TABLE IX. MULTITHREADING EFFICIENCY PERCENTAGE (> 50000)

Device	Th(x)= number of threads				
	Th-1	Th-2	Th-4	Th-8	Th-16
Device 1	0	0	1.00	0	0
Device 2	0	0	0.40	0.60	0
Device 3	0	0	1.00	0	0
Device 4	0	0	0.20	0.80	0
Device 5	0	0	1.00	0	0
Average	0	0	0.72	0.28	0

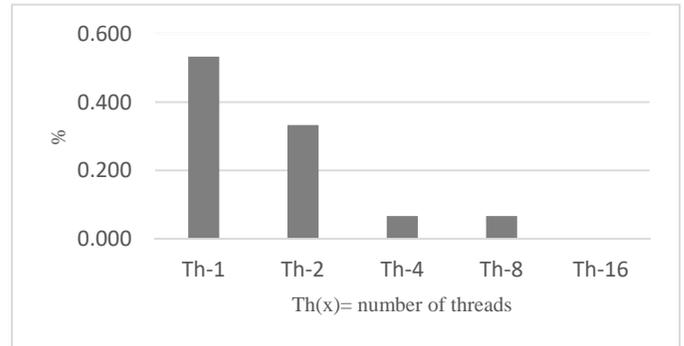


Fig. 8. Multithreading Efficiency Percentage (< 50000).

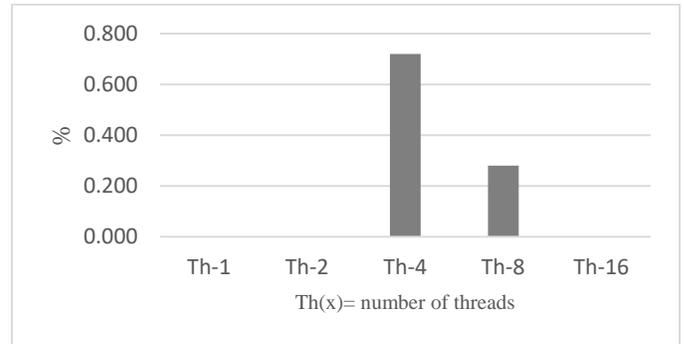


Fig. 9. Multithreading Efficiency Percentage (> 50000).

V. CONCLUSION

This study conducts an empirical experiment to determine the optimal number of threads to use in parallel merge sort. Several factors are discussed in this study to answer this question. First is the number of cores that impact multithreading performance. Second is the given data size that requires the use of multiple cores.

The implementation in this experiment takes a group of devices with various specifications. For each device, it takes fixed-sized data set and applies merge sort for sequential and parallel algorithms. For each device, the lowest average execution time (ms) is used to measure the efficiency of the experiment. Taking the average for all experiments, single-threaded is more efficient when the data size is less than 10^5 since it claimed 53%. Whereas, for data sizes exceeding 10^5 , multi-threaded implementation has better performance. The overall average of the experiments shows either four or eight threads are most efficient, with 72% and 28% respectively.

There were two main findings from these results. First, multithreading does not always have the most efficient runtime as it depends on the data size. Second, even when the data size increases, a specific number of threads will determine the optimized performance based on the device specifications. In other words, implementing as many threads as possible will not lead to higher runtime performance.

The conclusion is that if the data size is under 10^5 , single-threaded will be more efficient. In contrast, having multiple threads will perform better for data sizes that exceed 10^5 . In addition, the number of threads spawned should not exceed the number of cores (excluding merging threads).

REFERENCES

- [1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. "Introduction to algorithms," MIT press, 2009.
- [2] J. Katajainen, T. Pasanen, and J. Teuhola. "Practical in-place mergesort," Nord. J. Comput., 3(1):27–40, 1996.
- [3] M. Saadeh, H. Saadeh, and M. Qataweh, "Performance evaluation of parallel sorting algorithms on iman1 supercomputer," International Journal of Advanced Science and Technology, 95:57–72, 2016.
- [4] Merge Sort, howpublished = <https://www.geeksforgeeks.org/merge-sort/>, note = Accessed: 2021-12-01.
- [5] A. Uyar. "Parallel merge sort with double merging," In 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), pages 1–5. IEEE, 2014.
- [6] N. Parlante. "Linked list problems," Stanford CS Education Library, 1:33, 2002.
- [7] A. Abu Dalhoun, T. Kobbay, A. Sleit, M. Alfonseca, and A. Ortega. "Enhancing quicksort algorithm using a dynamic pivot selection technique," WULFENIA Journal, Austria, 19(10), 2012.
- [8] Z. Marszałek. "Parallelization of modified merge sort algorithm." Symmetry 9.9 : 176, 2017.
- [9] J. Holke, et al. "Data-adapted Parallel Merge Sort." European Conference on Parallel Processing. Springer, Cham, 2019.
- [10] D. P. Singh, Dharendra Pratap, I. Joshi, and J. Choudhary. "Survey of GPU based sorting algorithms." International Journal of Parallel Programming 46. : 1017-1034, 2018.
- [11] K. Raju, N. N. Chiplunkar, and K. Rajanikanth. "A CPU-GPU Cooperative Sorting Approach." 2019 Innovations in Power and Advanced Computing Technologies (i-PACT). Vol. 1. IEEE, 2019.
- [12] S. W. Hijazi, and M. Qataweh. "Study of Performance Evaluation of Binary Search on Merge Sorted Array Using Different Strategies." International Journal of Modern Education and Computer Science 9.12:1, 2017.
- [13] M. Jeon and D. Kim. "Parallel merge sort with load balancing," International Journal of Parallel Programming, 31(1):21–33, 2003.
- [14] K. B. Manwade. "Analysis of parallel merge sort algorithm," International Journal of Computer Applications, 1(19):66–69, 2010.
- [15] H. Rashid and K. Qureshi. "A practical performance comparison of parallel sorting algorithms on homogeneous network of workstations," WSEAS Transactions on Computers, 5(7):1606–1610, 2006.

Special Negative Database (SNDB) for Protecting Privacy in Big Data

Tamer Abdel Latif Ali¹, Mohamed Helmy Khafagy², Mohamed Hassan Farrag³

Computer Science Department, College of Computing & Information Technology¹

Arab Academy for Science, Technology and Maritime Transport, Aswan, Egypt¹

Computer Science Department, Faculty of Computers & Information, Fayoum University, Fayoum, Egypt²

Information Systems Department, Faculty of Computers & Information, Fayoum University, Fayoum, Egypt³

Abstract—Despite the importance of big data, it faces many challenges. The most important big data challenges are data storage, heterogeneity, inconsistency, timeliness, security, scalability, visualization, fault tolerance, and privacy. This paper concentrates on privacy which is one of the most pressing issues with big data. As mentioned in the Literature Review below there are numerous methods for safeguarding privacy with big data. This paper introduces an efficient technique called Specialized Negative Database (SNDB) for protecting privacy in big data. SNDB is proposed to avoid the drawbacks of all previous techniques. SNDB is based on deceiving bad users and hackers by replacing only sensitive attribute with its complement. Bad user cannot differentiate between the original data and the data after applying this technique.

Keywords—Big data; big data challenges; privacy violations; privacy-preserving techniques; special negative database; data integrity

I. INTRODUCTION

One of the most pressing challenges in big data is data privacy. Patients' data must be kept private since there is a risk of improper use of personal information being exposed when data from multiple sources is combined. In Privacy, every person has the right to select the extent of his or her interaction with the environment, as well as the amount of data that can be accessible by a third party. While it is sufficient to detect information as a "password" in security issues, since security is between two trusted parties, the server provider (SP) may be an adversary in privacy difficulties. We classified likely privacy violations in big data systems into four categories based on a literature review: data breaches, re-identification attacks, information gathering by service providers, and government tracking. The motivation of this manuscript is the importance of preserving privacy for everyone specially when dealing with big data. Also, the drawbacks of previous techniques like time consuming, losing data integrity, increasing size of data, low level of privacy and high complexity are one of the motivation factors for the author to propose a new technique called SNDB that will avoid drawbacks of other techniques. The next section will introduce literature review of previous techniques and their drawbacks. While in the third section, proposed technique and the manuscript contribution will be introduced. In fourth section, the author will introduce datasets used in proposed technique. Fifth section will discuss results and evaluation of the proposed technique when comparing with other

techniques. Finally, conclusion and future work will be introduced [1], [2], [3], [4].

II. LITERATURE REVIEW

A. Privacy Preserving by Slicing

Slicing is a method of dividing a dataset in vertical and horizontal manner. The process of dividing attributes into columns based on their correlations means vertical partitioning. Slicing can handle data with high dimensions according to attribute splitting. However, horizontal partitioning happens when records are combined into various buckets, and values in each column are permuted within each bucket randomly to disrupt the relationship between columns. The links between columns are broken by slicing, but the associations within each column are preserved [5].

B. Privacy in Big Data Generation Phase

1) *Access restriction*: Advertisement blockers, encryption methods, anti-tracking extensions, anti-virus software and anti-Malware are used to limit the access to sensitive data [6].

2) *Falsifying data*

- Socketpuppet is a deception-based method of masking an individual's internet identity [6].
- Users can use MaskMe to establish aliases for personal information such as their credit card number or email address [6].

C. Privacy in Big Data Storage Phase

1) *Attribute based encryption (ABE)*: ABE is a cloud storage encryption technique that assures big data privacy. The data owner defines the access policies in ABE, and data is encrypted according to those policies. Users whose features match with the data owner's access requirements can decrypt the encrypted data [6].

2) *Identity based encryption (IBE)*: IBE is used to simplify key management in a certificate-based public key infrastructure (PKI) by employing personal identities as public keys, such as an IP address or an email address, to maintain sender and receiver anonymity [6].

3) *Homomorphic encryption*: By calculating directly on the encryption of a message, it is possible to obtain the encryption of a function of that message [6].

4) *Storage path encryption*: The huge amount of data is first divided into numerous sequential parts, and each component is then saved on a distinct storage media controlled by several cloud storage providers [6].

5) *Usage of hybrid clouds*: The inherent qualities of public clouds, such as scalability and processing capacity, are combined with the inherent features of private clouds, such as security, to open up possible research opportunities in the processing and storage of enormous amounts of data [6].

D. Privacy in Big Data Processing Phase using Anonymization Techniques

1) *Generalization*: In the taxonomy of an attribute, a parent value is used to replace some values. An artist, rather than a singer or actor, might be used to symbolise a job attribute [6].

2) *Suppression*: A special character (e.g., *) is replaced for some values to declare that the modified value is not exposed in suppression. Value suppression, record suppression and cell suppression are examples of suppression schemes [6].

3) *Anatomization*: Rather of changing the quasi-identifier or sensitive features, anatomization separates the connection between the two [6].

4) *Permutation*: By dividing a set of data into groups and rearranging the sensitive values within each group, the connection between the quasi-identifier and the numerically sensitive feature is de-associated in permutation [6].

5) *Perturbation*: The actual data values are replaced with generated data values in perturbation, resulting in statistical information acquired from modified data that is statistically similar to that computed from the original data [6].

E. Privacy Protection Using Laws and Cyber Security

1) *Privacy laws and regulations*: Regulations and Laws have helped to protect privacy by limiting government tracking and limiting the reading, analysing, and publishing of users' personal information. Laws can also compel service providers to put in place necessary safeguards to protect data confidentiality and prevent data theft. This can enhance protecting privacy by avoiding privacy violations [7].

2) *Cyber security measures to prevent data breaches and cyber attacks* [7], [8].

- Honeypots and other espionage devices.
- Firewalls and other preventative measures.
- Malicious behavior is also detected via access logs and alert systems.
- Mechanisms for encrypting data.

F. Foggy Dummies

This approach is utilized in fog computing, and the fundamental idea is to create extremely intelligent dummies to preserve the user's privacy. This technique is used by the researcher to swap requests between fogs before sending them to server provider and then swapping the responses. This will be accomplished by fogs cooperating to exchange data before

sending it to the server provider [9].

G. Blind Third Party (BTP)

The essential point is why we must rely on a third party (TP) to keep the user safe from SP. That is, we are transferring the problem from one server to another. This strategy is dependent on fog's role as a middleman between the user and the SP in each location [9].

H. Double Foggy Cache

The primary idea behind this method is to use traditional cooperation to tackle the problem of peer trust. Furthermore, use SP to preserve your privacy. This strategy, in particular, can be seen as a significant advancement in the field. To accomplish this, we propose placing two caches in the Fog that will operate as intermediaries between peers. The first is for questions, while the second is for responses [9].

I. Secured Map Reduce Model (SMR)

As the data passes through the map-reduce phase, this new layer applies the security techniques to each individual piece of data. This security technique should be a simple encryption scheme, so that the complexity of new technique does not interfere with the big data's fundamental functioning. When data is processed using this suggested Secured Map Reduce (SMR) layer of big data, it can also be stored and secured. It begins with the collecting of data from social media, weblogs, and streaming data which is then delivered to Hadoop Distributed File System (HDFS). SMR is a suggested paradigm that adds a privacy layer between HDFS and the Map Reduce Layer (MR). Randomized procedures and perturbation were employed to strengthen the data's privacy [10].

J. Blind Peer Approach

This technique fixes the fundamental flaw in the prior technique, in which blind third party may collude with server provider to infringe on consumers' privacy. The new notion in the BLP strategy is to rely on collaboration with a large number of peers rather than dealing with a single TP. As a result of user's request would be sent to another peer in the same area, then encrypted by SPPK, giving the other peer no choice but to pass the question on to the SP, who would decrypt and resolve it [11].

K. Integrated Blind Parties (IBPs)

By integrating the BTP and BLP, This IBPs strategy raises the level of privacy while removing the disadvantages of the other seven options. When a peer isn't active in the area, the user can only rely on the BLP in this case. Furthermore, in the event of a resource shortage, without encrypting the query, the user might exchange it with another peer. In that circumstance, the peer can perform the BTP strategy rather than the user. This strategy can be used in any of the seven techniques [11].

L. Negative Database Conversion Algorithm

Instead of a single tuple, a negative database conversion technique is utilised to generate a big set of values. The data sets that have been generated are inserted into the database. In contrast to normal database applications, a harmful request in our negative database will be unable to access the database's

data. Because of the fabrication of fake sets of data in comparison to the premier data, the term negative is utilized. Both database encryption algorithms and virtual database encryption are used to encrypt the actual data [12].

M. Negative Database and Generic Database

The Entity, Attributes, and Values model of the general database design (EAV) was evaluated using the blob data storage type. The data collected, such as exam results, will be organized into three columns Entity, Attributes and Values as the name implies, the EAV is made up of three parts: entities, attributes, and values. The most straightforward approach to apply this principle is to create three tables for each data input (entity). There are two ways to implement the Negative database concept: one that statically generates negative data, i.e. the System Administrator defines the Negative data. Another that both statically and dynamically generates negative data. The user i.e. generates the dynamic negative data [13].

N. Enhanced BTP

In this enhanced approach, there is a new factor added to the old BTP, which is a unique token. This new technique consists of seven factors. A unique token is defined when the user sends a hidden code within a query to the service provider (SP) while SP returns the previous query token. Then SP will store the token for each ID generated by the third party, so the previous one cannot be used in a later query. When the third party inquiries from SP, a change will occur on the user's token, and the user will discover unauthorized access to his data by third party, so the proposed technique will be a powerful guarantee that there is no breakthrough [14].

O. Light Weight Cryptography Techniques(LWCT)

Based on an oil spill detection application, LWCT is utilized to secure a data transmission framework for the internet of things. Through locative and boundary value aggregation, this strategy eliminates duplicate data transmission. The suggested method protects data transfer by combining known lightweight cryptographic techniques with simple ID-based authentication [15], [16].

P. Block Nested Loop (BNL) Skyline Algorithms

This method is used to determine which encryption algorithm is best for ensuring data protection and privacy issue. The author of the Skyline algorithm considers two primary parameters: the rate of variation and the number of dimensions [17].

The author summaries all important drawbacks of previous techniques in the following factors:

- Time consuming
- High complexity
- Losing data integrity
- Low privacy protection
- Increasing size of data

The next section will introduce the major contribution of this manuscript. The author will propose a new technique based on negative database and the deception of bad users or hackers. The proposed technique is special negative database. This manuscript contribution can be summarized to enhance preserving privacy in big data and avoid time consuming, losing data integrity, complexity and violating any other big data challenges like its size, fault tolerance, timeliness and scalability.

III. PROPOSED TECHNIQUE

In this section, the authors introduce a new technique to protect privacy in big data in a new manner that deceives bad users or hackers.

Bad user cannot differentiate between the original data and the data after applying this technique. This technique is called Special Negative Database (SNDB). SNDB is based on deceiving bad users and hackers by replacing only sensitive attribute with its complement. SNDB takes into consideration all attribute types such as binomial, numeric, polynomial as mentioned in Fig. 1. The authors divided the technique into different cases as we will see in the following subsections.

A. Binomial

Binomial attribute means having only two values. Binomial consists of two categories one of them is binary and the other is Boolean. Binary consists of two values 0 or 1 but Boolean is like True or False values or other two values that are vice versa. SNDB deals with binomial attribute by replacing the value with its complement.

B. Numeric

The numeric attribute can be either integer or real numbers. SNDB deals with the numeric value in a different manner. It computes the complement of the digit into 9 individually with a maximum of 4 digits from right taking into consideration the national ID.

For example, if the numeric value is 2896547, the value after complement will be 2893452. While in real numbers, SNDB computes the complement of the digit into 9 individually with a maximum of 4 digits from left.

C. Date and Time

In the case of the date attribute, the date is divided into the year, month, and day. If the value represents a year, SNDB complement each value as a whole to current year when values are less than current year. If some values are equal to current year, the complement of each value will be a whole to Current year+1. If the value represents a month, SNDB will compute the complement of the value into 13. If the value represents a day, SNDB will compute the complement of the digit into 31.

In the case of time attribute. The time is divided into hours, minutes, and seconds. If the value represents hours, SNDB will compute the complement of the value into 24. If the value represents minutes or seconds, SNDB will compute the complement of the value into 9 for the right digit and into 6 for the left digit.

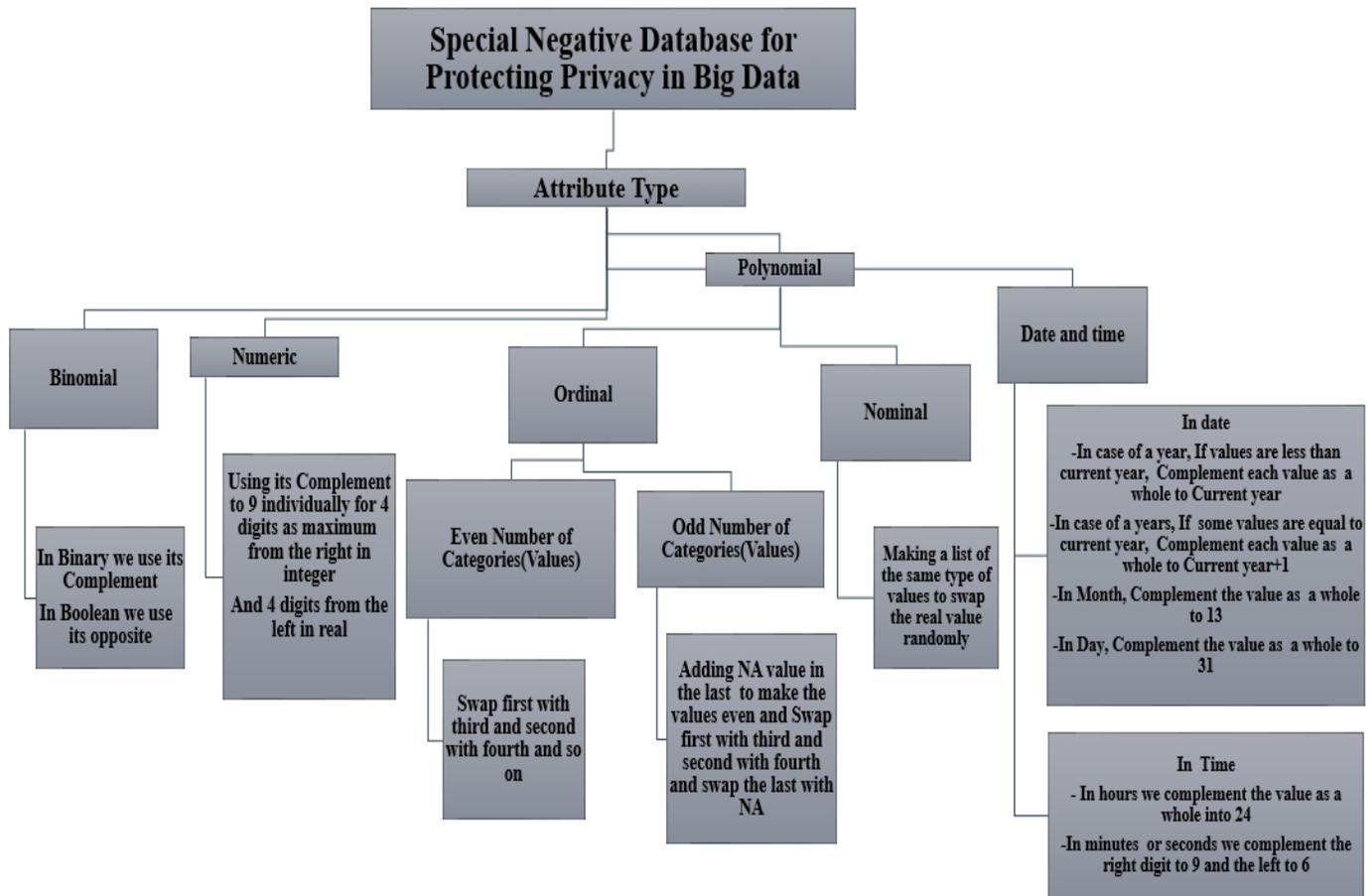


Fig. 1. Architecture of Special Negative Database (SNDB).

D. Polynomial

Polynomial means that the attribute has more than two text values. Polynomial is divided into two main types. The first is called ordinal and the second is called nominal.

1) *Ordinal*: Ordinal means that the values can be categorized, classified, ranked, and ordered. Drink size, for example, is an ordinal feature that correlates to the sizes of drinks available at fast-food restaurants. Small, medium, and large are the three possible values for this nominal attribute. These values have a logical order (which correlates to increasing drink size), but the values do not indicate how much larger a medium is than a large. Grade (e.g., A+, A, A-, B+, and so on) and professional rank are two further examples of ordinal characteristics [18].

In ordinal, the number of categories is very important in dealing with the swapping technique. Swapping is used to deceive the bad user that the data is real. In the case of an even number of categories, swapping is very easy to be implemented. SNDB will swap the first value with the third and the second with the fourth and so on. But in case of the odd number of categories, first, we will add the not allowed value (NA) at the last then swapping will be applied as an even manner.

2) *Nominal*: A nominal attribute's values are names of things or symbols. Nominal implies "related to names." Because each value reflects a state, code or category, nominal characteristics are also known as categorical. There is no discernible order to the values.

Enumerations are another term for values in computer science. Hair color and marital status are two attributes are examples of nominal. Black, blond, red, brown, auburn, grey, and white are all conceivable hair color values in our system. The value of single, married, divorced, or widowed can be assigned to the characteristic of marital status. Both marital status and hair color are also examples of nominal attributes. An occupation is another example of a nominal attribute, having values such as teacher, programmer, farmer dentist, and so on [18].

In the Nominal case, it's important to make a list of values to swap between the real value and one from this list randomly. If the values are names of persons, a list of names will be created. Then one value from this list will be selected and replaced randomly with the original one saving its index in the list. This operation will be repeated again and so on. Using the index, we can get the original data.

IV. DATASETS

In this paper, we apply our model on different datasets according to sensitive attributes taking into consideration all data types.

A. Pollution Dataset

This dataset is about pollution in the United States. The EPA has well-documented pollution in the United States, but downloading all of the data and arranging it in a format that data scientists are interested in is a nuisance. As a result, I gathered data for four key pollutants (nitrogen dioxide, carbon monoxide, ozone, and Sulphur dioxide) for every day between 2000 and 2016 and organized them in a CSV file. There are a total of twenty-eight fields. Each of the four pollutants (NO₂, CO, O₃, and SO₂) has its own set of five columns. 1746661 observations were made. The city, date local, and CO mean are all sensitive parameters.

B. Prouni

This dataset is about Brazil student's scholarship given by Brazilian government on the Prouni program. It contains data from 2005 to 2019 and each line of it corresponds to a student who benefits or has benefited from the Prouni program along with details about them. This dataset consists of 2692540 records.

V. RESULTS AND EVALUATIONS

This section will list the results and the evaluation of

applying SNDB technique on the datasets for privacy preserving. The author of this paper introduces results for applying SNDB on sensitive attributes in case of binomial, year date and full date and the rest of attribute types will be introduced in the next paper.

A. Binomial Results

Assuming that the sensitive attribute is BENEFICIARIO_DEFICIENTE_FISICO that means does student have special needs. SNDB swap the value nao that means no with the value sim that means yes and vice versa. Fig. 2 and Fig. 3 illustrate one sample of Prouni dataset before applying SNDB technique and another sample after applying it.

Fig. 4 shows computing some statistical operations on Prouni dataset before applying SNDB technique to get the type of scholarship according to special need. The results show that the number of students who have a special need and have got BOLSA PARCIAL 50% scholarship is 4,947 while the number of who do not have a special need with the same scholarship is 808,557. But students with special needs who have got BOLSA INTEGRAL scholarship is 14,222 while the number of the students who do not have special needs with the same scholarship is 1,862,484. On the other hand, the number of the students with special needs who have got BOLSA COMPLEMENTAR 25% scholarship is 4. While the number of who do not have special needs with the same scholarship is 2,326.

Row No.	ANO_...	CODIG...	NOME_JES...	TIPO_BOLSA	MODALIDAD...	NOME_CUR...	NOME_TUR...	SEX...	RACA_B...	DT_NASCIM...	BENEFICIA...	REGIAO_BE...	SIGLA_...	MUNICIPIO_...	idade
1	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Enfermagem	Integral	F	Branca	Feb 17, 1987	nao	SUL	RS	santo angelo	34
2	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Servico Social	Noturno	F	Parda	Jun 14, 1986	nao	SUL	RS	frederico wes...	35
3	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Servico Social	Noturno	F	Parda	Jun 3, 1984	nao	SUL	RS	frederico wes...	37
4	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Ciencia Da C...	Noturno	M	Branca	Oct 19, 1987	nao	SUL	RS	frederico wes...	33
5	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Ciencia Da C...	Noturno	M	Amarela	Jul 20, 1987	nao	SUL	RS	frederico wes...	34
6	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	M	Parda	Feb 13, 1985	nao	SUL	PR	sao jose dos ...	36
7	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	F	Branca	Jul 20, 1987	nao	SUL	PR	sao jose dos ...	34
8	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	F	Amarela	Sep 6, 1987	nao	SUL	PR	sao jose dos ...	34
9	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	F	Parda	Dec 27, 1987	nao	SUL	PR	sao jose dos ...	33
10	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	M	Branca	Apr 4, 1987	nao	SUL	PR	sao jose dos ...	34
11	2005	20	UNIVERSIDA...	BOLSA INTE...	PRESENCIAL	Educacao Fis...	Noturno	F	Branca	Jun 15, 1987	nao	SUL	RS	soledade	34
12	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Farmacia	Matutino	F	Branca	Feb 26, 1988	nao	SUL	RS	frederico wes...	33
13	2005	423	UNIVERSIDA...	BOLSA INTE...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Infor...	Jun 15, 1984	nao	SUL	RS	santiago	37
14	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Infor...	Sep 8, 1978	nao	SUL	RS	santiago	43
15	2005	423	UNIVERSIDA...	BOLSA INTE...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Infor...	Sep 22, 1976	nao	SUL	RS	santiago	45
16	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Infor...	Mar 15, 1957	nao	SUL	RS	santiago	64
17	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	F	Branca	May 11, 1985	nao	SUL	RS	santiago	36
18	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	M	Branca	Jan 26, 1983	nao	SUL	RS	santiago	38

Fig. 2. Sample of Prouni Dataset before Applying SNDB.

Row No.	ANO_CON...	CODIG...	NOME_IES...	TIPO_BOLSA	MODALIDAD...	NOME_CUR...	NOME_TUR...	SEX...	RACA_BENE...	DT_NASCIM...	BENE...	REGIA...	SIGL...	MUNICIPIO_BENEFIC...	idade
1	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Enfermagem	Integral	F	Branca	Feb 17, 1987	sim	SUL	RS	santo angelo	34
2	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Servico Social	Noturno	F	Parda	Jun 14, 1986	sim	SUL	RS	frederico westphalen	35
3	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Servico Social	Noturno	F	Parda	Jun 3, 1984	sim	SUL	RS	frederico westphalen	37
4	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Ciencia Da C...	Noturno	M	Branca	Oct 19, 1987	sim	SUL	RS	frederico westphalen	33
5	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Ciencia Da C...	Noturno	M	Amarela	Jul 20, 1987	sim	SUL	RS	frederico westphalen	34
6	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	M	Parda	Feb 13, 1985	sim	SUL	PR	sao jose dos pinhais	36
7	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	F	Branca	Jul 20, 1987	sim	SUL	PR	sao jose dos pinhais	34
8	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	F	Amarela	Sep 6, 1987	sim	SUL	PR	sao jose dos pinhais	34
9	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	F	Parda	Dec 27, 1987	sim	SUL	PR	sao jose dos pinhais	33
10	2005	10	PONTIFICIA...	BOLSA INTE...	PRESENCIAL	Administracao	Noturno	M	Branca	Apr 4, 1987	sim	SUL	PR	sao jose dos pinhais	34
11	2005	20	UNIVERSIDA...	BOLSA INTE...	PRESENCIAL	Educacao Fis...	Noturno	F	Branca	Jun 15, 1987	sim	SUL	RS	soledade	34
12	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Farmacia	Matutino	F	Branca	Feb 26, 1988	sim	SUL	RS	frederico westphalen	33
13	2005	423	UNIVERSIDA...	BOLSA INTE...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Informada	Jun 15, 1984	sim	SUL	RS	santiago	37
14	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Informada	Sep 8, 1978	sim	SUL	RS	santiago	43
15	2005	423	UNIVERSIDA...	BOLSA INTE...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Informada	Sep 22, 1976	sim	SUL	RS	santiago	45
16	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	M	Nao Informada	Mar 15, 1957	sim	SUL	RS	santiago	64
17	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	F	Branca	May 11, 1985	sim	SUL	RS	santiago	36
18	2005	423	UNIVERSIDA...	BOLSA PARC...	PRESENCIAL	Engenharia A...	Noturno	M	Branca	Jan 26, 1983	sim	SUL	RS	santiago	38

Fig. 3. Sample of Prouni Dataset after Applying SNDB.

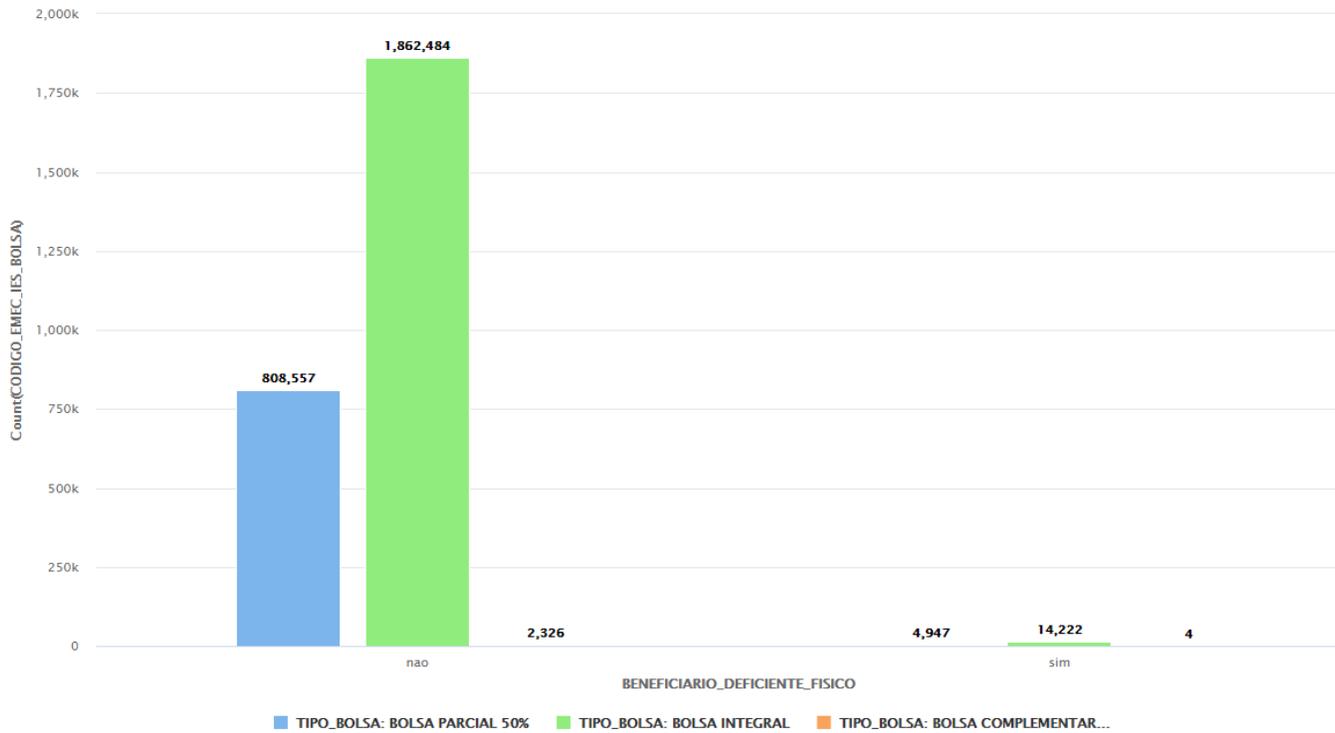


Fig. 4. Type of Scholarship with respect to Special need before Applying SNDB.

Fig. 5 shows computing some statistical operations on Prouni dataset after applying SNDB technique to get the type of scholarship according to special needs. The results show that the number of students who have a special need and have got BOLSA PARCIAL 50% scholarship is 808,557 while the number of who do not have special needs with the same scholarship is 4,947. But students with special needs who have got BOLSA INTEGRAL scholarship is 1,862,484 while the number of who do not have special needs with the same

scholarship is 14,222. On the other hand, the number of the students with special needs who have got BOLSA COMPLEMENTAR 25% scholarship is 2,326. While the number of students who do not have special needs with the same scholarship is 4. Fig. 4 and Fig. 5 show that the results have a big difference before applying SNDB and after applying SNDB on the same dataset which mean the success of the algorithm when applying on binomial sensitive attribute.

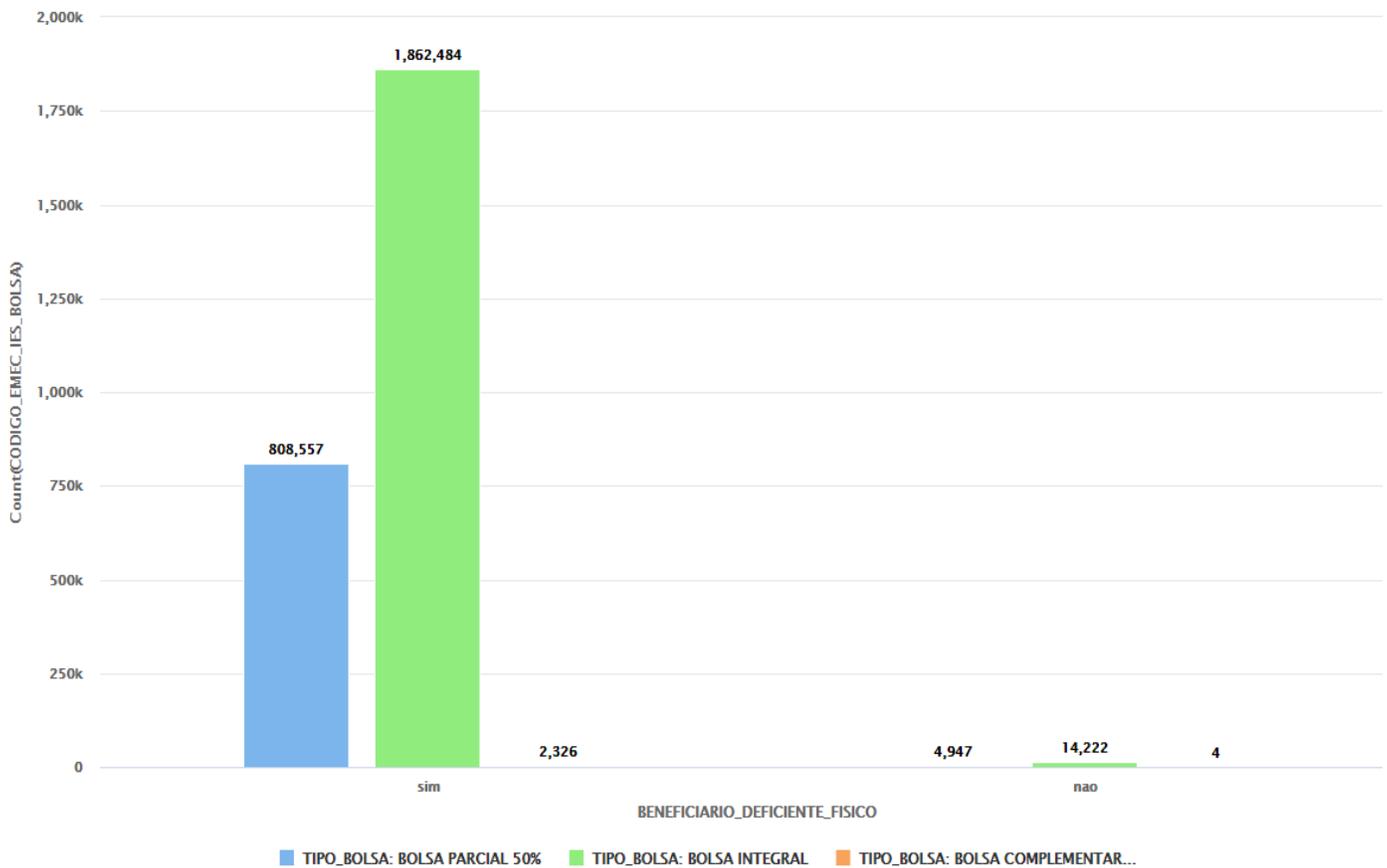


Fig. 5. Type of Scholarship with respect to Special Need after Applying SNDB.

B. Date Results

1) *Year date:* SNDB technique deals with sensitive attribute with the type date in a special manner. Assuming the sensitive attribute is ANO_CONCESSAO_BOLSA that means the year of the scholarship. It consists of a year only. SNDB applies the complement of a year to the current year (2021). Fig. 6 and Fig. 7 illustrate one sample of Prouni dataset before applying SNDB technique and another sample after applying SNDB on ANO_CONCESSAO_BOLSA.

Fig. 8 and Fig. 9 illustrate the difference between gender of the student who has got the scholarship before applying SNDB and after applying it. If we check out Fig. 8 and Fig. 9, we will say that the years of scholarship in the original data are from 2005 to 2019 while years from 2002 to 2016 are the years of scholarship after applying SNDB technique. When taking 2005 as an example, we will see the number of male students is 36,097 and the number of female students is 39,532 in the original data. While in the data after applying SNDB, the number of male students is 108,057 and the number of female students is 131,205 in original data. Fig. 8 and Fig. 9 show that the results have a big difference before applying SNDB and after applying SNDB on ANO_CONCESSAO_BOLSA in the same dataset which

means the success of the algorithm when applying on date sensitive attribute of year value only. In the next section the paper will show the result of applying SNDB on different date value.

Full date: Assuming the sensitive attribute is date local. SNDB deals with date local in a different manner, SNDB changes the month only because changing the month is sufficient to change the original data. SNDB swaps January with December and vice versa, February with November and vice versa, March with October and vice versa, April with September and vice versa, May with August and vice versa and Jun with July and vice versa. Fig. 10 and Fig. 11 illustrate one sample of pollution dataset before applying SNDB technique and another sample after applying SNDB on date local attribute. Fig. 12 and Fig. 13 below illustrate the difference between the maximum value of (Sulphur Dioxide and Nitrogen Dioxide) mean before and after applying SNDB on date local. This paper takes the first five days of December,2000 as an example to show the difference between the original dataset and the dataset after applying SNDB technique. When checking out Fig. 12 and 13, we will see the big difference between the values of original and SNDB dataset.

Row No.	ANO_CO...	CODIG...	NOME_IES_BOLSA	TIPO_BOLSA	MODALIDAD...	NOME_CUR...	NOME...	SEX...	RACA_BENE...	DT_NASCIM...	BE...	REGIAO_BE...	SIGLA...	MUNICIPIO_...	idade
1198348	2013	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	F	Branca	Dec 19, 1994	nao	SUDESTE	SP	macaubal	26
1198349	2013	146	CENTRO UNIVERSI...	BOLSA INTE...	PRESENCIAL	Agronomia	Noturno	F	Branca	Jul 7, 1990	nao	SUDESTE	MG	barbacena	31
1198350	2013	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	F	Branca	Feb 4, 1995	nao	SUDESTE	MG	itapagipe	26
1198351	2013	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	F	Branca	Nov 27, 1995	nao	SUDESTE	MG	frutal	25
1198352	2013	146	CENTRO UNIVERSI...	BOLSA INTE...	PRESENCIAL	Agronomia	Noturno	F	Branca	Oct 26, 1995	nao	SUDESTE	SP	guapiacu	25
1198353	2013	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	M	Branca	Aug 8, 1995	nao	SUDESTE	SP	avare	26
1198354	2013	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Biomedicina	Noturno	F	Parda	Nov 16, 1995	nao	SUDESTE	MG	frutal	25
1198355	2013	338	PONTIFICIA UNIVER...	BOLSA PARC...	PRESENCIAL	Engenharia ...	Matutino	F	Branca	Sep 23, 1995	nao	SUDESTE	SP	jose bonifacio	26
1198356	2013	338	PONTIFICIA UNIVER...	BOLSA INTE...	PRESENCIAL	Engenharia ...	Matutino	M	Parda	Nov 30, 1992	nao	SUDESTE	MG	campos gerais	28
1198357	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	EAD	Pedagogia	A Dista...	F	Branca	Feb 1, 1984	nao	SUDESTE	RJ	sao goncalo	37
1198358	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Feb 16, 1989	nao	SUDESTE	RJ	sao goncalo	32
1198359	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Jul 3, 1981	nao	SUDESTE	RJ	rio de janeiro	40
1198360	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Jun 12, 1994	nao	SUDESTE	RJ	volta redonda	27
1198361	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Branca	Oct 31, 1994	nao	SUDESTE	RJ	barra mansa	26
1198362	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Branca	Sep 16, 1993	nao	SUDESTE	RJ	quatis	28
1198363	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Branca	Apr 21, 1992	nao	SUDESTE	MG	santa rita de j...	29
1198364	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Jul 21, 1993	nao	SUDESTE	RJ	barra mansa	28
1198365	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	EAD	Administracao	A Dista...	F	Parda	Mar 14, 1993	nao	SUDESTE	RJ	mesquita	28
1198366	2013	163	UNIVERSIDADE ES...	BOLSA INTE...	EAD	Marketing	A Dista...	M	Parda	Sep 4, 1991	nao	SUDESTE	RJ	belford roxo	30

Fig. 6. Sample of Prouni Dataset before Applying SNDB on ano_concessao_bolsa.

Row No.	ANO_CO...	CODIG...	NOME_IES_BOLSA	TIPO_BOLSA	MODALIDAD...	NOME_CUR...	NOME...	SEX...	RACA_BENE...	DT_NASCIM...	BE...	REGIAO_BE...	SIGLA_...	MUNICIPIO_...	idade
1198348	2008	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	F	Branca	Dec 19, 1994	nao	SUDESTE	SP	macaubal	26
1198349	2008	146	CENTRO UNIVERSI...	BOLSA INTE...	PRESENCIAL	Agronomia	Noturno	F	Branca	Jul 7, 1990	nao	SUDESTE	MG	barbacena	31
1198350	2008	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	F	Branca	Feb 4, 1995	nao	SUDESTE	MG	itapagipe	26
1198351	2008	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	F	Branca	Nov 27, 1995	nao	SUDESTE	MG	frutal	25
1198352	2008	146	CENTRO UNIVERSI...	BOLSA INTE...	PRESENCIAL	Agronomia	Noturno	F	Branca	Oct 26, 1995	nao	SUDESTE	SP	guapiacu	25
1198353	2008	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Agronomia	Noturno	M	Branca	Aug 8, 1995	nao	SUDESTE	SP	avare	26
1198354	2008	146	CENTRO UNIVERSI...	BOLSA PARC...	PRESENCIAL	Biomedicina	Noturno	F	Parda	Nov 16, 1995	nao	SUDESTE	MG	frutal	25
1198355	2008	338	PONTIFICIA UNIVER...	BOLSA PARC...	PRESENCIAL	Engenharia ...	Matutino	F	Branca	Sep 23, 1995	nao	SUDESTE	SP	jose bonifacio	26
1198356	2008	338	PONTIFICIA UNIVER...	BOLSA INTE...	PRESENCIAL	Engenharia ...	Matutino	M	Parda	Nov 30, 1992	nao	SUDESTE	MG	campos gerais	28
1198357	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	EAD	Pedagogia	A Dist...	F	Branca	Feb 1, 1984	nao	SUDESTE	RJ	sao goncalo	37
1198358	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Feb 16, 1989	nao	SUDESTE	RJ	sao goncalo	32
1198359	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Jul 3, 1981	nao	SUDESTE	RJ	rio de janeiro	40
1198360	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Jun 12, 1994	nao	SUDESTE	RJ	volta redonda	27
1198361	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Branca	Oct 31, 1994	nao	SUDESTE	RJ	barra mansa	26
1198362	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Branca	Sep 16, 1993	nao	SUDESTE	RJ	quatis	28
1198363	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Branca	Apr 21, 1992	nao	SUDESTE	MG	santa rita de j...	29
1198364	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	PRESENCIAL	Enfermagem	Noturno	F	Parda	Jul 21, 1993	nao	SUDESTE	RJ	barra mansa	28
1198365	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	EAD	Administracao	A Dist...	F	Parda	Mar 14, 1993	nao	SUDESTE	RJ	mesquita	28
1198366	2008	163	UNIVERSIDADE ES...	BOLSA INTE...	EAD	Marketing	A Dist...	M	Parda	Sep 4, 1991	nao	SUDESTE	RJ	belford roxo	30

Fig. 7. Sample of Prouni Dataset after Applying SNDB on ano_concessao_bolsa.

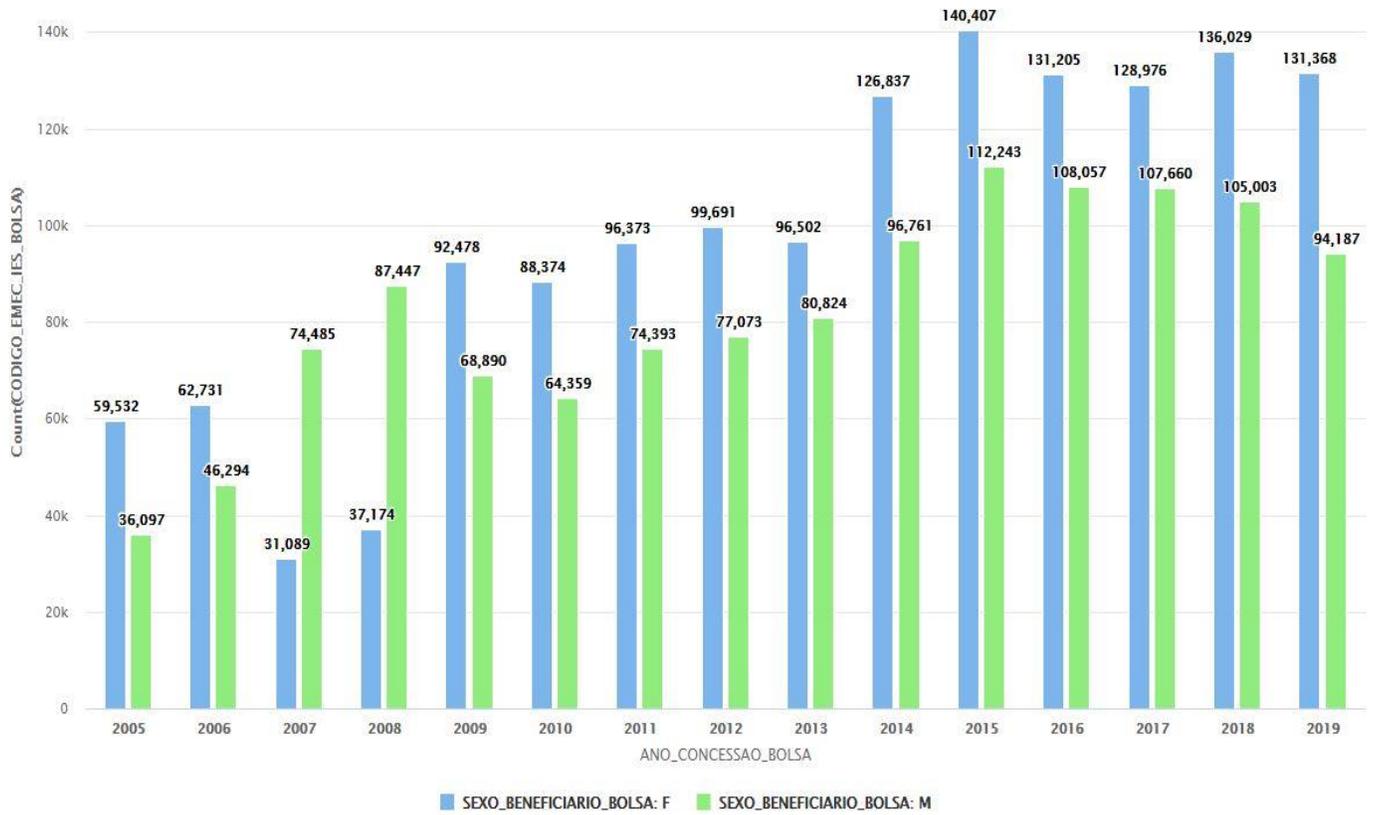


Fig. 8. Gender of Students in the Year of Scholarship before Applying SNDB.

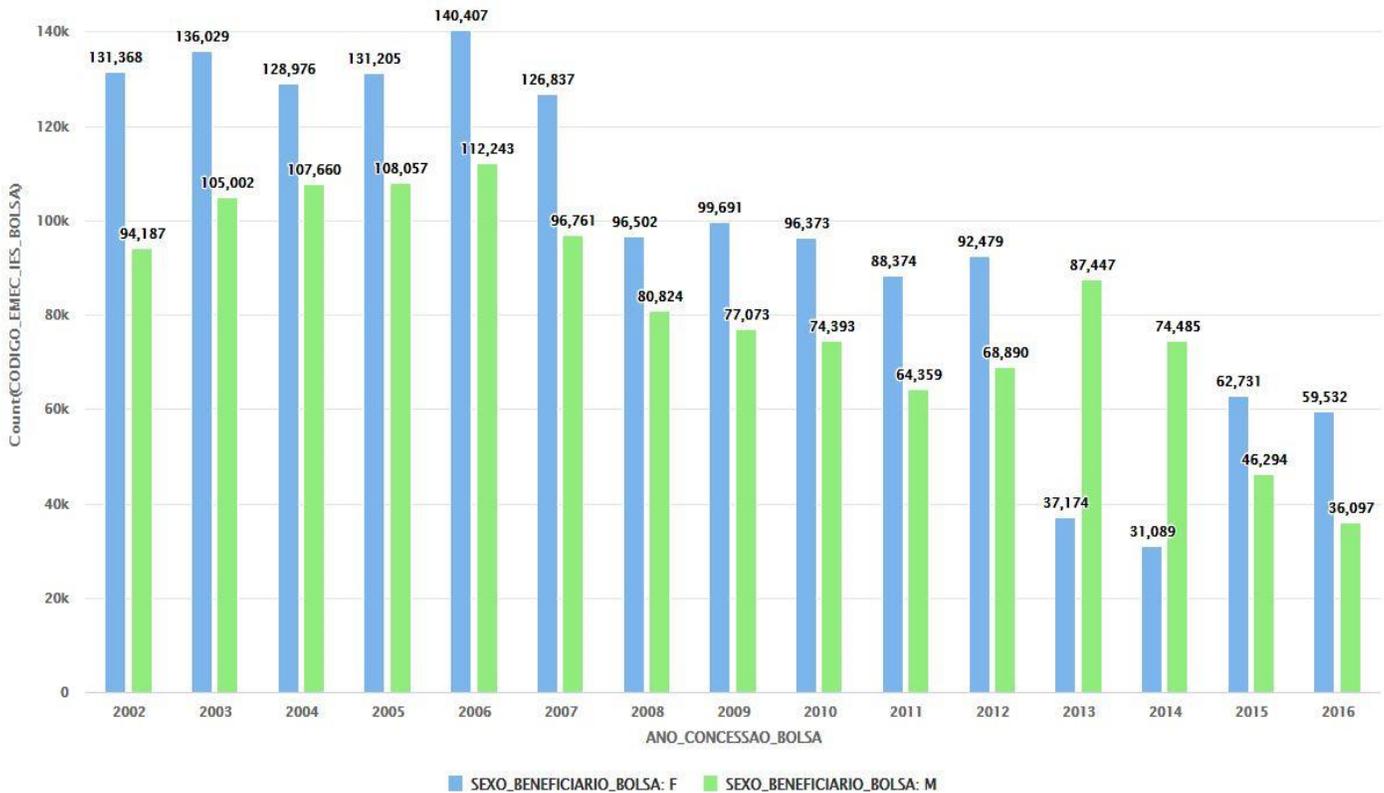


Fig. 9. Gender of Students in the Year of Scholarship after Applying SNDB.

Row ...	State Code	County Code	Site Num	Address	State	County	City	Date Local	NO2 Units	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour
1	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 1, 2000	Parts per billion	19.042	49	19
2	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 1, 2000	Parts per billion	19.042	49	19
3	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 1, 2000	Parts per billion	19.042	49	19
4	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 1, 2000	Parts per billion	19.042	49	19
5	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 2, 2000	Parts per billion	22.958	36	19
6	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 2, 2000	Parts per billion	22.958	36	19
7	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 2, 2000	Parts per billion	22.958	36	19
8	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 2, 2000	Parts per billion	22.958	36	19
9	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 3, 2000	Parts per billion	38.125	51	8
10	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 3, 2000	Parts per billion	38.125	51	8
11	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 3, 2000	Parts per billion	38.125	51	8
12	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 3, 2000	Parts per billion	38.125	51	8
13	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 4, 2000	Parts per billion	40.261	74	8
14	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 4, 2000	Parts per billion	40.261	74	8
15	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 4, 2000	Parts per billion	40.261	74	8
16	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 4, 2000	Parts per billion	40.261	74	8
17	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 5, 2000	Parts per billion	48.450	61	22
18	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Jan 5, 2000	Parts per billion	48.450	61	22

Fig. 10. Sample of Pollution Dataset before Applying SNDB on Date Local.

Row ...	State Code	County Code	Site Num	Address	State	County	City	Date Local	NO2 Units	NO2 Me...	NO2 1st Max Value	NO2 1st Max Hour
1	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 1, 2000	Parts per billion	19.042	49	19
2	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 1, 2000	Parts per billion	19.042	49	19
3	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 1, 2000	Parts per billion	19.042	49	19
4	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 1, 2000	Parts per billion	19.042	49	19
5	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 2, 2000	Parts per billion	22.958	36	19
6	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 2, 2000	Parts per billion	22.958	36	19
7	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 2, 2000	Parts per billion	22.958	36	19
8	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 2, 2000	Parts per billion	22.958	36	19
9	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 3, 2000	Parts per billion	38.125	51	8
10	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 3, 2000	Parts per billion	38.125	51	8
11	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 3, 2000	Parts per billion	38.125	51	8
12	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 3, 2000	Parts per billion	38.125	51	8
13	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 4, 2000	Parts per billion	40.261	74	8
14	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 4, 2000	Parts per billion	40.261	74	8
15	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 4, 2000	Parts per billion	40.261	74	8
16	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 4, 2000	Parts per billion	40.261	74	8
17	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 5, 2000	Parts per billion	48.450	61	22
18	4	13	3002	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	Dec 5, 2000	Parts per billion	48.450	61	22

Fig. 11. Sample of Pollution Dataset after Applying SNDB on Date Local.

As seen from results figures above, SNDB is valid for big data since the size of data does not change before and after applying SNDB. Processing time is fast when comparing to traditional negative database. The deception of SNDB technique is big and this makes privacy level is stronger. Bad users or hackers cannot differentiate between the original data and the data after applying SNDB technique. This makes the

decryption very hard for bad users while it is very easy for data owner to decrypt the SNDB data. Table I below shows the comparison between SNDB and traditional negative database. There are comparative results on different datasets because SNDB technique is applied according to sensitive attributes covering all data types of each dataset.

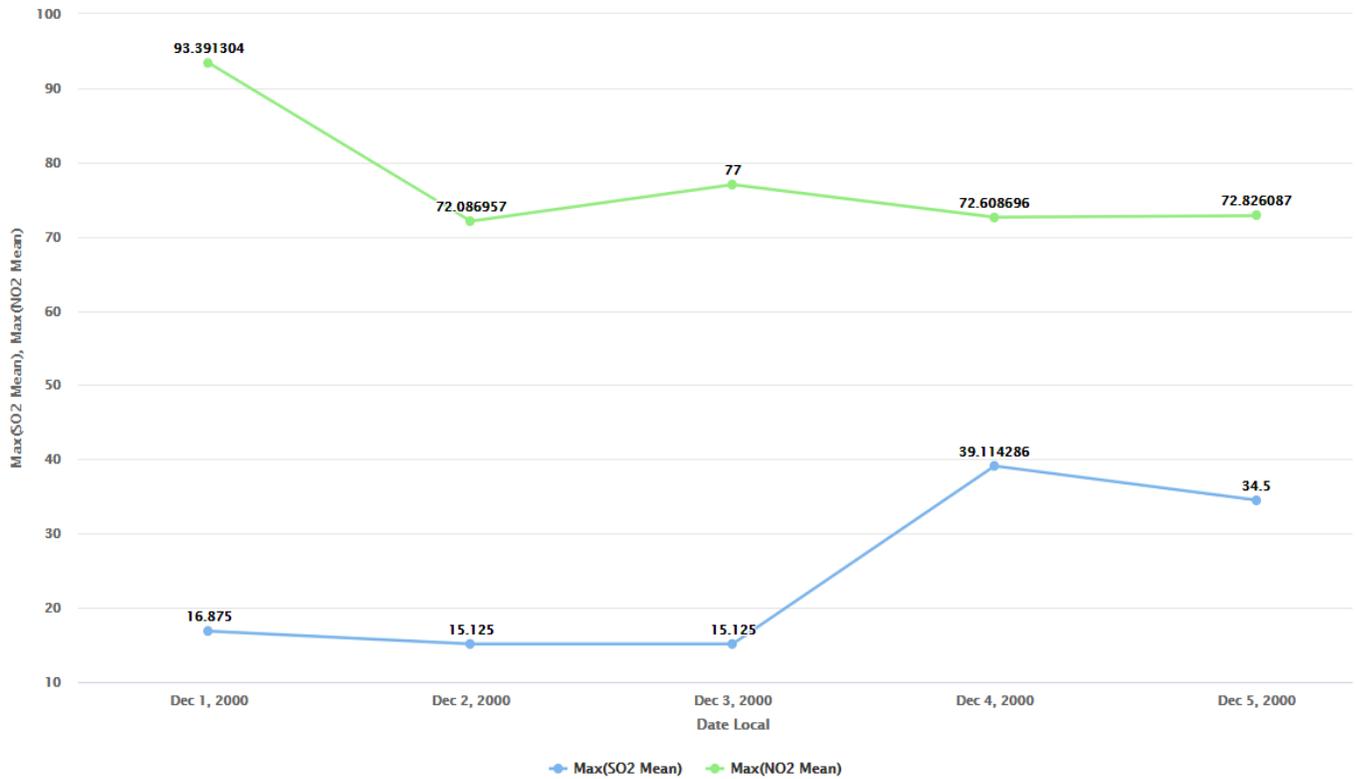


Fig. 12. Maximum Value of (Sulphur Dioxide and Nitrogen Dioxide) mean before Applying SNDB on Date Local.

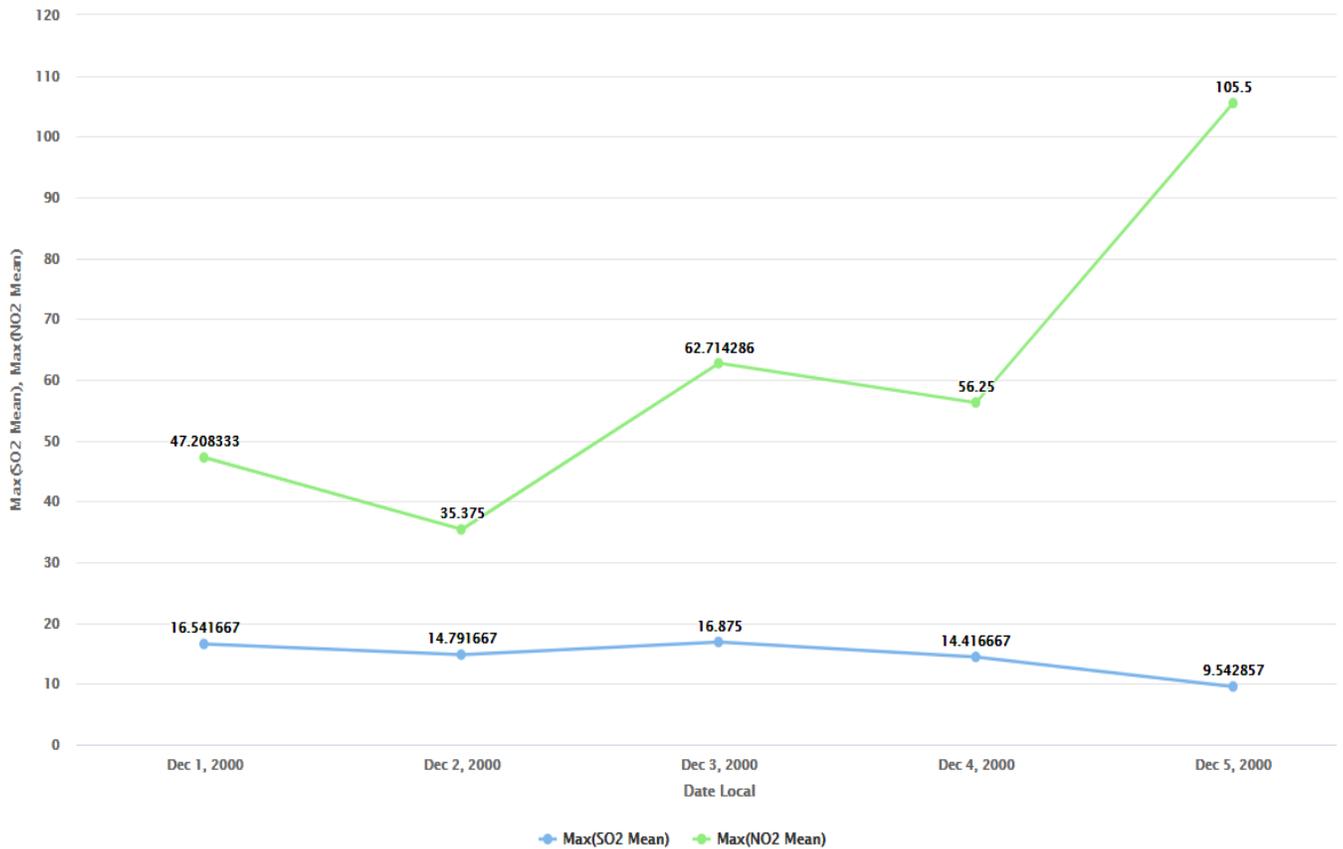


Fig. 13. Maximum Value of (Sulphur Dioxide and Nitrogen Dioxide) mean after Applying SNDB on Date Local.

TABLE I. COMPARISON BETWEEN TRADITIONAL NEGATIVE DATABASE AND SNDB

Technique	Records in the original pollution dataset	Records in the pollution dataset after applying the algorithm	Records in the original Prouni dataset	Records in the Prouni dataset after applying the algorithm	Processing Time	Deception level	Privacy level	Decryption for users and hackers	Decryption for data owner	Validity for Big data
Traditional Negative Database	1,746,661	3,493,322 In case of full date attribute	2,692,540	5,385,080 In case of binomial and year date attributes	Slow 32 seconds in case of Pollution dataset and 47 seconds in case of Prouni dataset	There is a little deception for bad users and hackers	Strong	Hard	Easy	Not valid because the size of data increases and doubles
SNDB	1,746,661	1,746,661 In case of full date attribute	2,692,540	2,692,540 In case of binomial and year date attributes	Fast 9 seconds in case of Pollution dataset and 12 seconds in case of Prouni dataset	There is a big deception for bad users and hackers	More stronger	Harder	Easier	Valid because the size of data does not change

VI. CONCLUSION

This paper lists the most important big data challenges and focuses on privacy challenge; it summaries privacy violation situations. The author also provides a list of the most efficient and popular techniques used to protect data privacy with their advantages and drawbacks. The proposed technique in this paper is SNDB based on negative database in different manner. SNDB is based on deceiving bad users and hackers by replacing only sensitive attribute with its complement. SNDB takes into consideration all attribute types such as binomial, numeric, polynomial. SNDB technique is applied on different datasets according to the type of the sensitive attributes of each dataset. In this technique, bad user cannot differentiate between the original data and the data after applying this technique which enhances the level of privacy.

As seen from results, SNDB can avoid drawbacks of previous techniques since it has the advantage of high privacy protection in big data. SNDB has no time consuming since it deals with sensitive attribute only. It also keeps track of data integrity and data size since there is no decreasing or increasing for any record of data and this advantage makes SNDB very suitable for big data. It also has low complexity since it only replaces sensitive attribute value with its complement. After applying SNDB, we can easily get the original data by applying the complement another time according to the rules of the data owner.

VII. FUTURE WORK

Finally, the author provides the results of applying SNDB on big dataset with binomial, year date and full date sensitive attribute. In the future work, the author will introduce the results of applying SNDB on numeric, ordinal and nominal sensitive attributes. Also, the author tends to take into consideration transposition techniques instead of replacing values with each other's.

REFERENCES

- [1] P. V. Desai, "A survey on big data applications and challenges," In Proc. of the Second International Conf. on Inventive Communication and Computational Technologies (ICICCT), IEEE, pp. 737-740, 2018.
- [2] M. M. Shendi, H. M. Elkadi, and M. H. Khafagy, "A study on the big data log analysis: goals, challenges, issues, and tools," International Journal of Artificial Intelligence and Soft Computing, vol. 7, no. 2, pp. 5-12, 2019.
- [3] Sk. M. Gouse and G. K. Mohan, "Improving the Performance of Various Privacy Preserving Databases using Hybrid Geometric Data Perturbation Classification Model," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 10, pp. 249-253, 2020.
- [4] O. Almutairi and K. Almarhabi, "Investigation of Smart Home Security and Privacy: Consumer Perception in Saudi Arabia," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 4, pp. 614-622, 2021.
- [5] K. Vani and B. Srinivas, "Enhanced slicing for privacy-preserving data publishing," The International Journal of Engineering and Science (IJES), vol. 2, no. 10, pp. 1-4, 2013.
- [6] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo "Protection of big data privacy," IEEE access, vol. 4, pp. 821-1834, 2016.
- [7] J. A. Shamsi and M. A. Khojaye, "Understanding privacy violations in big data systems," IT Professional, vol. 20, no. 3, pp. 73-81, 2018.
- [8] L. Rajesh and P. Satyanarayana, "Detecting Flooding Attacks in Communication Protocol of Industrial Control Systems," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 1, pp. 396-401, 2020.
- [9] A. A. Abi Sen, F. A. Eassa, and K. Jambi, "Preserving privacy of smart cities based on the fog computing," In Proc. International Conf. on Smart Cities, Infrastructure, Technologies and Applications, Springer, Cham, pp. 185-191, 2017.
- [10] P. Jain, M. Gyanchandani and N. Khare, "Enhanced secured map-reduce layer for big data privacy and security," Journal of Big Data, vol. 6, no. 1, pp. 1-17, 2019.
- [11] M. Yamin, Y. Alsaawy, A. B. Alkhodre, and A. A. A. Sen, "An innovative method for preserving privacy in internet of things," Journal of Sensors, vol. 19, no. 9, pp. 3355, 2019. [Online]. Available: <https://doi.org/10.3390/s19153355>.
- [12] A. Patel, N. Sharma, and M. Eirinaki, "Negative Database for Data Security," In Proc. International Conf. on Computing, Engineering and Information, IEEE, pp. 67-70, 2009.

- [13] C. Egbunike and S. Rajendran, "The Implementation of Negative Database as a Security Technique on a Generic Database System," In Proc. International Conf. on circuits Power and Computing Technologies (ICCPCT), IEEE, pp. 1-8, 2017.
- [14] A. A. A. Sen, F. A. Eassa, K. Jambi, N. M. Bahbouh, S. S. Albouq, and A. Alshantqi, "Enhanced-blind approach for privacy protection of iot," 2020 IEEE 7th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, pp. 240-243, 2020.
- [15] S. AR and B. G. Banik, "A Comprehensive Study of Blockchain Services: Future of Cryptography," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 11, no. 10, pp. 279-288, 2020.
- [16] H. V. Abhijith and H. S. Rameshbabu, "Secure data transmission framework for internet of things based on oil spill detection application," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 5, pp. 189-195, 2021.
- [17] S. Trichni, F. Omary, and M. Bougrine, "New smart encryption approach based on multidimensional analysis tools," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 5, pp. 666-675, 2021.
- [18] H. Jiawei, M. Kamber, and P. Jian, *Data Mining Concepts and Techniques*, (3rd ed), Morgan Kauffman, 2011.

Drug Sentiment Analysis using Machine Learning Classifiers

Mohammed Nazim Uddin^{1*}, Md. Ferdous Bin Hafiz²
Sohrab Hossain³

Department of Computer Science and Engineering
School of Science, Engineering and Technology
East Delta University, Chattogram, Bangladesh

Shah Mohammad Mominul Islam⁴

Department of Electrical and Electronic Engineering
School of Science, Engineering and Technology
East Delta University
Chattogram, Bangladesh

Abstract—In recent times, one of the most emerging sub-dimensions of natural language processing is sentiment analysis which refers to analyzing opinion on a particular subject from plain text. Drug sentiment analysis has become very significant in present times as classifying medicines based on their effectiveness through analyzing reviews from users can assist potential future consumers in gaining knowledge and making better decisions about a particular drug. The objective of this proposed research is to measure the effectiveness level of a particular drug. Currently most of the text mining researches are based on unsupervised machine learning methods to cluster data. When supervised learning methods are used for text mining, the usual primary concern is to classify the data into two classes. Lack of technical terms in similar datasets make the categorization even more challenging. The proposed research focuses on finding out the keywords through tokenization and lemmatization so that better accuracy can be achieved for categorizing the drugs based on their effectiveness using different algorithms. Such categorization can be instrumental for treating illness as well as improve one's health and well-being. Four machine learning algorithms have been applied for binary classification and one for multiclass classification on the drug review dataset acquired from the UCI machine learning repository. The machine learning algorithms used for binary classification are naive Bayes classifier, random forest, support vector classifier (SVC), and multilayer perceptron; among these machine learning algorithms, linear SVC was used for multiclass classification. Results obtained from these four classifier algorithms have been analyzed to evaluate their performances. The random forest has been proven to have the best performance among these four algorithms. However, multiclass classification was found to have low performance when applied to natural language processing. On the contrary, the applied linear SVC algorithm performed better for class 2 with AUC 0.82 in this research.

Keywords—Machine Learning Algorithms; natural language processing; drugs sentiment analysis; text mining

I. INTRODUCTION

Among many research dimensions in Natural Language Processing (NLP), sentiment analysis has become one of the promising fields of research in the recent century [1] [2]. A wide range of research domains has been covered by sentiment analysis, i.e. economy, polity, and medicine. In the pharmaceutical industry, large volumes of online user's views are evaluated automatically in order to obtain useful information about the efficacy and side effects of

pharmaceuticals that could be utilized to enhance pharmacovigilance systems. Throughout the years, Sentiment analysis techniques have grown significantly in the last decade, evolving from basic rules to advanced machine learning techniques like deep learning, which has become a prominent technology in many NLP tasks. This triumph is not lost on sentiment analysis. Besides, several machine learning systems have recently been shown to be better than previous methods. These methods have achieved impactful results on standard sentiment analysis datasets [3] [4].

Various aspects such as 'medical condition', 'treatment procedure', etc., use the medical sentiment as a research study that directly impacts the users' health conditions. Any sort of progress or deterioration can be identified by analyzing patient status periodically. The medical condition can be expressed implicitly or explicitly. Mentioning the symptoms is a part of the implicit sentiment in the medical context. For example, consider the statement: 'I recently started Lexapro 3 days, I'm on extreme weight losses'. The term "weight losses here do not reflect a negative sense; however, it implies the negative medication side effect, where sentiment is defined as negative in the preceding statement. Hence, for making correct interpretations, additional information is required.

On the contrary, analyzing the health conditions is relatively much easier in the case of explicit sentiment. For instance, considering the statement, "I recently started Lexapro 3 days, I'm absolutely lost I feel weak and shaky every day". The words absolutely 'lost', 'weak', and 'shaky' are used to describe symptoms in this statement. Deciding about patients' medical issues is an important aspect, specifically when they learn from other patients' experiences, i.e., choosing a hospital, clinic, and medication [5]. Hospitals gain from this information since it allows them to understand better and address the interests and concerns of their patients. The experience covered with sentiment analysis and passions are being shared by the patients; sentiment analysis is being taught by the power of this type of experiment since this type of study identifies people's sentiment about a topic as well as its characteristics. The medical material available on the internet is completely free. Manually analyzing such a large volume of data is ineffective because of its existence in large volumes. Assessed examinations are denoted as positive or negative, for the most part, based on the pre-programmed acceptance of extreme suppositions. The online and traditional

*Corresponding Author.

review methods are supplanted by notion investigation nowadays, which is led by organizations for finding a broad conclusion regarding their products and service. As a result, their marketing approach and product awareness increase, and user management improve. It is quite imperative to be broken down since a tremendous amount of content is available online. That includes deep comprehensions of standard dialects are included in the programmed examination of this data. In our everyday life, thoughts and sentiments play an essential role. Basic leadership, learning, correspondence, and mindfulness in human circumstances are assisted. Socially produced regional substances are becoming prevalent in online life; hence the importance of dealing with and comprehending vernacular content is growing. Existing materials notwithstanding, such as nearby sayings, Myths, and fables are unearthed, widely disseminated on the internet.

Compared to reviews of other products, drug reviews are investigated less. When analyzed, drug reviews are primarily utilized to categorize a particular drug as a positive or a negative one as multi-class classification from text mining can be unyielding. The proposed research facilitates the categorization of drugs not into two categories rather into five classes based on their effectiveness. Such an outcome can be beneficial for both consumers and manufacturers to understand the effectiveness of drugs as well as whether a particular drug has any significant side-effect.

This paper is organized as follows: Section II illustrates literature review for sentiment analysis, Section III explains the algorithms those classify the sentiment of the drugs, Section IV portrays the findings of the study, and Section V demonstrates the contributions of the proposed research.

II. LITERATURE REVIEW

Supervised machine learning methods used by Twitter datasets, such as support bigram, vector machines and unigram, were analyzed by a research study led by Balahur (2013) [6]. Following the applications of these approaches to Twitter data, the results indicated that methods of unigram and bigram support vector machines are outshined. Emotive words, modifiers and unique tags were included in these results, enhancing the performance rating of emotions. Another study conducted by Jianqiang et al. [7] (2018) presented an approach that is word embedded using unsupervised learning as a base. This suggested technique makes use of hidden contextual semantic connections and characterization between words and tweets. The

characteristics of mood polarity and n-gram are combined with the score of the embedded word to structure. A deep convolutional neural network was used to include a collection of emotional characteristics. Facebook, Instagram, and Twitter are a few examples of social media platform that helps to generate data and circulate content quickly. The amount of hate utterances has risen significantly while circulating content related to a particular topic. To filter these sorts of utterances, a research study presented by Schmidt and Wiegand [8] suggested a filtering tool for natural language processing. According to the results, character-level strategies are superior to token-level ones. The authors' methodology demonstrated that using a lexical list of resources to rank them might be beneficial when utilized with others. Based on K-means and cuckoo searching methods, Pandey et al. [9] suggested a unique metaheuristic approach. The best feasible cluster heads are found using this method based on the Twitter dataset's emotional subject material. Wang and Li [10] categorized the changed text algorithms to anticipate motions in image data for the sentiment analysis. Textual and visual features for labelling emotions inside an image are unsuitable for the forecast, according to their technique. The authors conducted experiments on two datasets and found that the recommended technique outperforms current methods in terms of accuracy. Unique research on Hierarchical Deep Fusion (HDF) emotional analysis methodology was studied by Xu et al. [11]. The relationship between the properties of text, images and sentimental content has been analyzed in the proposed model. The authors combined visual content with textual content using three-level Hierarchical Long Short Term Memory (H-LSTM) to investigate the inter-modal association of text and image at various levels. Some of the most widely applied machine learning and deep learning algorithms have been described in Table I.

Most of the above-mentioned researches primarily focus on unsupervised learning method. Compared to other product reviews, number of researches conducted on drug reviews is significantly low. One of the key challenges with dataset similar to the one used in this research is the lack of technical terms. A few researches that utilize supervised learning methods, perform binary classification as multi-class classification using the existing machine learning algorithms have been proven to be challenging. In this research, tokenization and lemmatization identify the key words. Also multi-class classification has been performed unlike the researches mentioned in this section.

TABLE I. EXITING ALGORITHMS ACCURACIES

Algorithm	Description
Neural Network	The neural network approach technique has a very high performance. It is a widely used technique for sentiment analysis and is capable of detecting all possible interactions between attributes. It is effective for dealing with a nonlinear connection between variables that is complex. The main disadvantage is that it takes longer to compute than other algorithms [12].
Naive Bayes	It is a Bayes' theorem based probabilistic classifier. Researchers use this method less commonly to make a prediction. The primary advantage is that it is scalable in comparison to other algorithms [13].
Support Vector machine	It is also a way of supervised machine learning for classification and regression analysis. When dealing with small datasets, the Support Vector Machine is very effective. It is more efficient than other approaches of classification and regression [14].
Decision Tree	Decision trees are simple but extensively used tools for prediction. IF-THEN rules can be simply converted from a decision tree. According to a previous study, prediction and forecast can be done using a decision tree. It can predict drug sentiment with low accuracy [15].
K-Nearest Neighbor	It is a popular pattern recognition method that is less non-parametric. It has the ability to utilize both regression and classification. It provides the finest performance and precision. It is the most basic machine learning algorithm [16].
AdaBoost	AdaBoost combines a number of poor classifiers to create a powerful one by iteratively retraining and weighing the classifiers depending on their accuracy [17].
Logistic regression	The log odds of the dichotomous result can be modelled as a linear combination of the predictor factors using this method [18].
Convolutional neural network	A feed-forward neural network that has been trained to extract key characteristics for the prediction job at hand. Nonlinear functions are used to filter features via convolutions. The dimensionality can then be reduced via pooling [19].
Maximum entropy	The greatest entropy concept is used to create a probabilistic classifier [20].
Conditional random fields	Given an observation series based on a conditional probability distribution across label sequences, this approach for segmenting and labelling structured data can be used [21].

III. METHODOLOGY

The dataset that used in this experiment has been collected from the UCI repository [22]. Tokenization and lemmatization were performed on the data after collecting the dataset. Four machine learning algorithms have been applied to the dataset for binary classification. The classes for binary classification are class 0, and class 1, where class 0 represents the effective drugs and class 1 indicates the ineffective drugs. The algorithms used for binary classification are naïve Bayes classifier, random forest (RF), support vector classifier (SVC) and multilayer perceptron (MLP). Linear SVC has also been applied to the dataset for multiclass classification.

Table II represents the classes for multiclass classification along with counts for each class. Class 0 represents highly effective drugs with a count of 1741, whereas class 1 represents considerably effective drugs with a count of 1238. Class 2 represents moderately effective drugs with a count of 529. In addition, class 3 represents marginally effective drugs with a count of 329. Besides, class 4 represents ineffective drugs with a count of 263.

A. Tokenization

In the case of natural language processing, a series of well-defined processes need to be carried out for analyzing the text. One of the primary processes is known as tokenization which plays a crucial role in the efficiency and correctness of the entire analysis. Tokenization refers to splitting the text into meaningful smaller units known as tokens. In most cases, tokens are identified as words or word sequences. Tokens are usually recognized when a white space character is encountered just after scanning a token. Preprocessing of text for punctuation removal and uppercase to lowercase conversion are often involved with the tokenization process [23].

B. Lemmatization

One of the most essential and elementary processes associated with natural language processing (NLP) is lemmatization. The base or dictionary form of a word is called a lemma. The term lemmatization refers to the morphological conversion of a word that exists in the textual form of the dataset to its lemma. The basic idea for this conversion is the removal of the declension from the end part of the word. In the case of a verb, lemma represents the infinitive form; in the case of a noun, lemma represents the singular form, and in the case of adjective or adverb, lemma represents the positive form. For example, the lemma for the word 'better' is 'good'. The lemma for the word 'brought' is 'bring'. Lemmatization can be perceived as a normalization method in which various morphological variants of a word are analyzed as a single item by mapping them into the same underlying lemma. As the aggregate number of specific terms are reduced, the complexity for analyzing the text is significantly decreased and thus, the overall time and resource utilization is improved. Lemmatization is widely applied for preprocessing the text in information retrieval, document clustering, sentiment analysis, etc. [24].

Table III shows tokenization and lemmatization of the feature benefitsReview.

TABLE II. DATA STATISTICS FOR CLASSIFICATION

Class	Effectiveness	Count
0	Highly Effective	1741
1	Considerably Effective	1238
2	Moderately Effective	572
3	Marginally Effective	329
4	Ineffective	263

After simplifying equation 3, we get the following probabilistic model:

$$P(X|Y_1, Y_2, \dots, Y_n) \propto P(X) * \prod_{i=1}^n P(Y_i|X) \quad (4)$$

Equation 4 exhibits the probabilistic model, which is obtained after simplifying equation 3.

From this probabilistic model, a classifier model is generated by calculating the probability of all given inputs for the possible values of the class variable, and the maximum value is identified to determine the specific value of the class. It can be expressed as follows:

$$X = \text{argmax} P(X) * \prod_{i=1}^n P(Y_i|X) \quad (5)$$

2) *Random forest*: The high variance problem of the decision tree classifier, where a minor change in the training data set can produce a very different tree, makes the decision tree classifier unstable. To eradicate this issue, the concept of the random forest was proposed which is an ensemble of decision trees. Random forest is a classifier with various classification methods or a single method with various parameters from the dataset. Assume a learning data set $D = ((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ that consists of n vectors, where $x \in X$ (X is a set of numerical observations) and $y \in Y$ (Y is a set of class labels). For a classification instance, a classifier maps $X \rightarrow Y$. Each tree of the forest is responsible for classifying a new input vector. Random forest combines the idea of bootstrapping data from a learning dataset to form training data set and selecting parameters randomly to construct decision trees. Bootstrap refers to selecting three-fourths of the learning dataset (sometimes two-thirds of the learning dataset) and replacing the rest of the data with some of the selected samples. While constructing a decision tree, features and their positions as nodes in a particular tree are chosen randomly. Thus random forest classifier, h , can be defined as:

$$h = \{h_1(X), h_2(X), \dots, h_n(X)\} \quad (6)$$

In equation 6, h_k is a decision tree having parameters θ_k , which is a subset of features chosen randomly [27].

3) *Support vector classifier*: Support Vector Classifier is a supervised learning algorithm to analyze data for classification. One of the distinctive properties of this algorithm is the ability to reduce empirical classification error and expand the geometric margin simultaneously. SVC provides high accuracy in text categorization, image classification, hand-written digits recognition, data classification, sentiment analysis, etc. SVC constructs a maximal separating hyperplane. Two parallel hyperplanes, which represent two different classes, are built on both sides of this separating hyperplane. By using SVC, an input vector is mapped to any of these two parallel hyperplanes. The purpose of the separating hyperplane is to maximize the space between the two parallel hyperplanes. It is assumed that less classification error can be achieved with a higher distance or margin between the parallel hyperplanes.

Consider a training dataset $((x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ where each x_n is a p dimensional vector that maps to corresponding y_n , which specifies the class. The value of y_n can be either -1 or $+1$. The equation of the separating hyperplane can be written as:

$$w \cdot x - b = 0 \quad (7)$$

In equation 7, w is a one-dimensional vector and b is a scalar. The equations for the two parallel hyperplanes can be written in equation 8 and equation 9 as follows:

$$w \cdot x + b = 1 \quad (8)$$

$$w \cdot x + b = -1 \quad (9)$$

The distance between these two hyperplanes is $\frac{b}{|w|}$ which implies that by minimizing $|w|$, we can maximize the distance between the two hyperplanes and thus achieve high performance.

In the case of hard margin where no misclassification is allowed from the training dataset, the problem can be stated as:

Minimize $|w|$ for

$$y_i(w^T x_i - b) \geq 1 \text{ for } i = 1, 2, \dots, n \quad (10)$$

In equation 10, the classifier is determined by solving w and b for the above problem statement.

In the case of soft margin where a few misclassifications are allowed from the training data set to achieve better accuracy for testing dataset, the problem can be stated as:

Minimize $|w|$ for

$$\left[\frac{1}{n} \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i - b)) \right] + \lambda |w|^2 \quad (11)$$

In equation 11, the trade-off between putting x_i in the right hyperplane and maximizing the distance between the parallel hyperplanes is determined by λ [28].

4) *Multilayer perceptron*: Inspired by the functioning procedure of human nervous systems, the concept of artificial neural networks has been developed and applied to design mathematical models to solve complex classification or regression problems. The building blocks of the artificial neural network are 'artificial neurons' or 'neurons'. Frequently these neurons are referred to as nodes. In a multilayer perceptron, which is a feedforward artificial neural network, these neurons are organized in layers and completely interconnected with each other via edges to construct a directed graph. The term 'feedforward' refers that this graph as acyclic. Each of these edges is associated with a real number which is called the weight of the edge. The layers of multilayer perceptron neural networks are the input layer, a number of hidden layers, and the output layer. For each neuron, there exists a summation function and an activation function. The summation function can be written as:

$$S_j = \sum_{i=1}^n w_{ij} I_i + \beta_j \quad (12)$$

In equation 12,

w_{ij} = The connection weight from neuron I_i to I_j

β_j = Bias weight

n = The total number of neuron inputs

The output of this summation function becomes an input of the activation function. There are various types of activation functions. One of the most applied activation functions is a nonlinear 'S' shaped curved sigmoid activation that can be expressed as:

$$f(x) = \frac{1}{1+e^{-x}} \quad (13)$$

Applying this activation function from equation 13, the output of the neuron I_j can be expressed as equation 14, which is shown below:

$$y_j = f_j(\sum_{i=1}^n w_{ij} I_i + \beta_j) \quad (14)$$

Once the neural network is constructed, the set of weights are tuned to estimate the required result [29].

IV. RESULTS

A confusion matrix has been developed as a binary prediction for each algorithm utilized to evaluate the performance system. One of the widely utilized methods for calculating predictions is a binary prediction which consists of the most significant building blocks of a ROC curve [30]. Each classification problem contains two classes. There exist two sets of positive and negative ((P) and (N)) labels of class for every instance. There are four possible categories for a classifier instance. True positive (TP) refers to the number of positive instances being classified appropriately. Similarly, true negative (TN) represents the number of negative instances being classified without any error. Opposite to that, if a positive instance is classified as a negative instance, it is considered as false positive (FP). Likewise, if a negative example is classified as a positive example, it is labelled as a false negative. While applying the algorithms to the dataset, in each instance, 80 percent of the dataset has been utilized as training data and the rest 20 percent has been used for testing.

We found accuracy, precision (P), recall (R), and F1 score by using the following equations:

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (15)$$

In equation 15, accuracy is measured as the total number of correctly identified cases divided by the total number of test cases.

$$\text{Precision (P)} = TP / (TP + FP) \quad (16)$$

In equation 16, precision is measured as number of true positive cases divided by number of all predicted positive cases.

$$\text{Recall (R)} = TP / (TP + FN) \quad (17)$$

In equation 16, recall is measured as the number of true positive cases divided by all actual positive cases.

$$\text{F1 Score} = 2 * (R * P) / (R + P) \quad (18)$$

In equation 18, F-1 score is measured from the calculated precision and recall values.

Table IV shows the confusion matrix for the proposed machine learning classifiers for all features. Random forest works as the best classifier with 94% accuracy among the four algorithms. The accuracies for MLP, SVC, and NB are almost the same. Later we calculate the accuracy by changing the number of features ranging from 5000 to 9000.

Table V shows the accuracy for the proposed machine learning classifiers for features ranging from 5000 to 9000. The proposed model provides consistent results for different features ranging from 5000 to 9000.

Fig. 3 shows ROC curves for individual classes and all classes for linear SVC. Fig. 3(a) shows that the area under the ROC curve is 0.56, which indicates that linear SVC is not very efficient in the case of identifying class 0 (highly effective drugs). Fig. 3(b) shows that the area under the ROC curve is 0.64, which indicates that linear SVC is slightly more efficient in the case of identifying class 1 (considerably effective drugs) than class 0. Fig. 3(c) shows that the area under the ROC curve is 0.82, which indicates that linear SVC is very efficient in the case of identifying class 2 (moderately effective drugs). Fig. 3(d) shows that the area under the ROC curve is 0.65, which indicates that linear SVC is approximately as efficient in the case of identifying class 3 (marginally effective) as class 1. Fig. 3(e) shows that the area under the ROC curve is 0.59, which indicates that linear SVC is slightly less efficient in the case of identifying class 4 (ineffective) than class 3. Fig. 3(f) plots precision against recall for all the classes. After analyzing the ROC curve, it is conspicuous that linear SVC has a significant positive performance in identifying class 2 drugs.

TABLE IV. ACCURACY FOR 5000, 6000, 7000, 8000 AND 9000 FEATURES

RF		MLP		SVC		NB	
938	16	829	49	872	6	857	21
46	36	75	83	113	45	96	62

TABLE V. ACCURACY FOR 5000, 6000, 7000, 8000 AND 9000 FEATURES

features	RF	MLP	SVC	NB
5000	94.21	88.71	88.80	88.71
6000	94.02	87.55	88.90	88.42
7000	93.92	84.75	88.51	88.71
8000	94.02	84.75	88.51	88.42
9000	94.11	88.32	88.42	88.61

TABLE VI. CLASSIFICATION REPORT FOR RF

	precision	recall	f1-score	support
Class 0	0.69	0.44	0.54	82
Class 1	0.95	0.98	0.97	954
accuracy	0.94			1036
macro avg	0.82	0.71	0.75	1036
weighted avg	0.93	0.94	0.93	1036

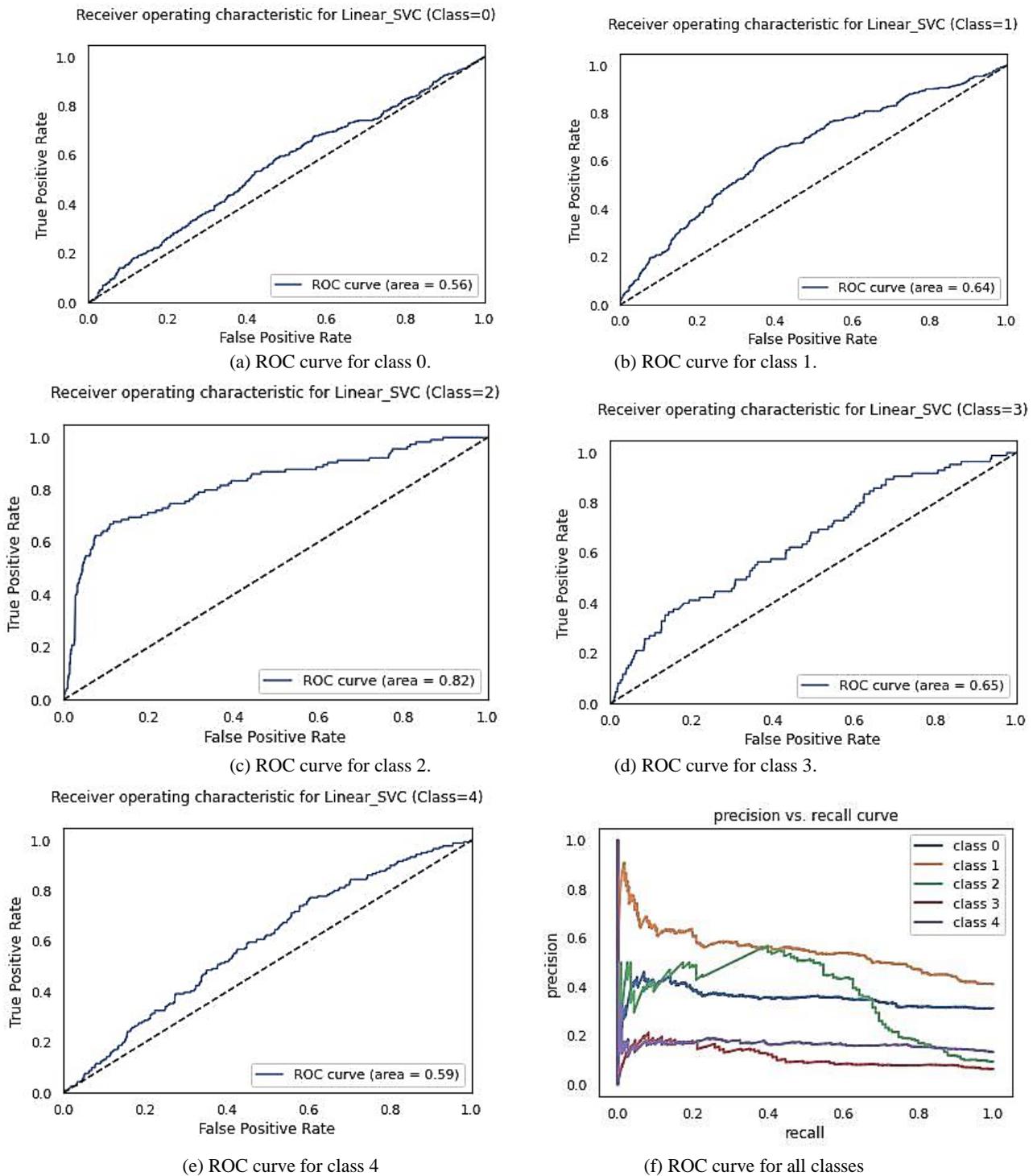


Fig. 3. ROC Curve for Linear SVC.

The evaluation metric used to represent the multiple class classification is the ROC curve which plots True Positive Rate (TPR) against False Positive Rate (FPR) at different threshold settings. Though the ROC curve is applicable for binary classification, there is an alternative way to integrate it for multiple-class classification. This approach is known as 'one vs rest' where a multiple-class problem is treated as a binary classification problem. In a ROC curve, a higher value in the

X-axis indicates a greater false positive rate than the true negative rate. A higher value in the Y-axis indicates a greater false-negative rate than the true positive rate. The efficient way to discriminate between accurate and inaccurate classification is to measure the area under the ROC curve (AUC). This area accepts values between 0 and 1, where 0 indicates the completely inaccurate classification of the class and 1 indicates perfectly accurate classification. Although

multiclass sentiment classification is extremely challenging for textual data, but Fig. 3(c) shows very promising accuracy in this research.

V. CONCLUSION

From the experimental result, the calculated average accuracy for Radom Forest, Multilayer Perceptron, Support Vector Classifier and Naïve Bayes Classifier is 94.06%, 86.82%, 88.63% and 88.57%, respectively. It is found that the Random Forest algorithm has generated the best accuracy among the four algorithms. In the case of Random Forest, higher precision, recall, and f1-score have been achieved for effective drugs compared to those measurements of ineffective drugs. The reason behind calculating the f-1 score is to get accuracy measurement from a different perspective as the f-1 score delivers the balance between precision and recall. Although multiclass classification is a challenging task for sentiment analysis, linear SVC shows the promising result for class 2 (moderately effective drugs). In this research, we have applied five machine learning algorithms. Unlike most of the similar researches in NLP when text mining is used for clustering the data, supervised learning methods have been implemented in this research to gain a better understanding of a drug by measuring its level of effectiveness. It can play an important role for curing diseases. In future, we intend to apply deep learning algorithms like Long Short Term Memory Networks (LSTMs), Generative Adversarial Networks (GANs), Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs). In addition to that, we would like to implement multi-language sentiment analysis using data-driven approaches.

REFERENCES

- [1] M. Levis, C. L. Westgate, J. Gui, B. V. Watts, and B. Shiner, "Natural language processing of clinical mental health notes may add predictive value to existing suicide risk models," (in English), *Psychological Medicine*, Article vol. 51, no. 8, pp. 1382-1391, Jun 2021, Art. no. Pii s0033291720000173.
- [2] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, "Natural Language Processing Reveals Vulnerable Mental Health Support Groups and Heightened Health Anxiety on Reddit During COVID-19: Observational Study," (in English), *Journal of Medical Internet Research*, Article vol. 22, no. 10, p. 16, Oct 2020, Art. no. e22635.
- [3] M. Lewis, R. Li, M. Booth, M. Latymer, C. Borlenghi, and M. Kissner, "Sentiment analysis of Altmetric data using a human-augmented, artificial intelligence (AI) approach," (in English), *Current Medical Research and Opinion, Meeting Abstract* vol. 37, pp. 16-16, Apr 2021.
- [4] T. H. McCoy, V. M. Castro, A. Cagan, A. M. Roberson, I. S. Kohane, and R. H. Perlis, "Sentiment Measured in Hospital Discharge Notes Is Associated with Readmission and Mortality Risk: An Electronic Health Record Study," (in English), *Plos One*, Article vol. 10, no. 8, p. 10, Aug 2015, Art. no. e0136341.
- [5] A. Sansone et al., "The Sentiment Analysis of Tweets as a New Tool to Measure Public Perception of Male Erectile and Ejaculatory Dysfunctions," (in English), *Sexual Medicine*, Article vol. 7, no. 4, pp. 464-471, Dec 2019.
- [6] A. Balahur, "Sentiment analysis," in 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2013, pp. 120-128.
- [7] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," *IEEE Access*, vol. 6, pp. 23253-23260, 2018.
- [8] A. Schmidt and M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 2017, pp. 1-10.
- [9] A. C. Pandey, D. S. Rajpoot, and M. Saraswat, "Data clustering using hybrid improved cuckoo search method," in 2016 Ninth International Conference on Contemporary Computing (IC3), 2016, pp. 1-6.
- [10] Y. Wang and B. Li, "Sentiment Analysis for Social Media Images," in 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 2015, pp. 1584-1591.
- [11] J. Xu et al., "Sentiment analysis of social images via hierarchical deep fusion of content and links," *Appl. Soft Comput.*, vol. 80, pp. 387-399, Jul 2019.
- [12] I. S. Alimova and E. V. Tutubalina, "Entity-Level Classification of Adverse Drug Reaction: A Comparative Analysis of Neural Network Models," *Programming and Computer Software*, vol. 45, no. 8, pp. 439-447, Dec 2019.
- [13] M. Devaraj, R. Piryani, and V. K. Singh, "Lexicon Ensemble and Lexicon Pooling for Sentiment Polarity Detection," *Iete Technical Review*, vol. 33, no. 3, pp. 332-340, May-Jun 2016.
- [14] S. S. Liu and I. Lee, "Email Sentiment Analysis Through k-Means Labeling and Support Vector Machine Classification," *Cybernetics and Systems*, vol. 49, no. 3, pp. 181-199, 2018.
- [15] S. Amin, M. I. Uddin, D. H. AlSaeed, A. Khan, and M. Adnan, "Early Detection of Seasonal Outbreaks from Twitter Data Using Machine Learning Approaches," *Complexity*, vol. 2021, Mar 2021, Art. no. 5520366.
- [16] X. Y. Wang, D. Li, M. Y. Jiang, Z. L. Pei, and L. Xu, "K Nearest Neighbor Algorithm Coupled with Metabonomics to Study the Therapeutic Mechanism of Sendeng-4 in Adjuvant-Induced Rheumatoid Arthritis Rat," *Evidence-Based Complementary and Alternative Medicine*, vol. 2018, 2018, Art. no. 2484912.
- [17] A. Tharwat, T. Gaber, A. E. Hassaniien, and M. Elhoseny, "Automated toxicity test model based on a bio-inspired technique and AdaBoost classifier," *Computers & Electrical Engineering*, vol. 71, pp. 346-358, Oct 2018.
- [18] Y. Choi and Y. Boo, "Comparing Logistic Regression Models with Alternative Machine Learning Methods to Predict the Risk of Drug Intoxication Mortality," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, Feb 2020, Art. no. 897.
- [19] K. Yanagisawa et al., "Convolutional Neural Network Can Recognize Drug Resistance of Single Cancer Cells," *International Journal of Molecular Sciences*, vol. 21, no. 9, May 2020, Art. no. 3166.
- [20] K. Yanagisawa et al., "Convolutional Neural Network Can Recognize Drug Resistance of Single Cancer Cells," *International Journal of Molecular Sciences*, vol. 21, no. 9, May 2020, Art. no. 3166.
- [21] K. Y. Chang, T. P. Lin, L. Y. Shih, and C. K. Wang, "Analysis and Prediction of the Critical Regions of Antimicrobial Peptides Based on Conditional Random Fields," *Plos One*, vol. 10, no. 3, Mar 2015, Art. no. e0119490.
- [22] Archive.ics.uci.edu. S. Kallumadi and F. Gräßer 2018. UCI Machine Learning Repository: Drug Review Dataset (Druglib.com) Data Set. [online] Available at: <<https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29>> [Accessed 24 October 2021].
- [23] Mullen et al., (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open Source Software*, 3(23), 655, <https://doi.org/10.21105/joss.00655>
- [24] A. Ozcift, K. Akarsu, F. Yumuk, and C. Soylemez, "Advancing natural language processing (NLP) applications of morphologically rich languages with bidirectional encoder representations from transformers (BERT): an empirical case study for Turkish," *Automatika*, vol. 62, no. 2, pp. 226-238, Apr 2021.
- [25] F. H. Khan, U. Qamar, and S. Bashir, "Senti-CS: Building a lexical resource for sentiment analysis using subjective feature selection and normalized Chi-Square-based feature weight generation," *Expert Systems*, vol. 33, no. 5, pp. 489-500, Oct 2016.
- [26] M. Devaraj, R. Piryani, and V. K. Singh, "Lexicon Ensemble and Lexicon Pooling for Sentiment Polarity Detection," *Iete Technical Review*, vol. 33, no. 3, pp. 332-340, May-Jun 2016.

- [27] Azar, Ahmad Taher et al. "A random forest classifier for lymph diseases." *Computer methods and programs in biomedicine* vol. 113,2 (2014): 465-73. doi:10.1016/j.cmpb.2013.11.004
- [28] M. H. Afif and A. Hedar, "Data classification using Support Vector Machine integrated with scatter search method," 2012 Japan-Egypt Conference on Electronics, Communications and Computers, 2012, pp. 168-172, doi: 10.1109/JEC-ECC.2012.6186977.
- [29] Basheer I, Hajmeer M (2000) Artificial neural networks: fundamentals, computing, design, and application. *J Microbiol Methods* 43(1):3–31
- [30] F. Kunneman, M. Lambooi, A. Wong, A. van den Bosch, and L. Mollema, "Monitoring stance towards vaccination in twitter messages," *Bmc Medical Informatics and Decision Making*, vol. 20, no. 1, Feb 2020, Art. no. 33.

Effective Malware Detection using Shapely Boosting Algorithm

Rajesh Kumar, Geetha S
School of Computer Science and Engineering
Vellore Institute of Technology
Chennai, India

Abstract—Malware constitutes a prime exploitation tool to attack the vulnerabilities in software that lead to a threat to security. The number of malware gets generated as exploitation tools need effective methods to detect them. Machine learning methods are effective in detecting malware. The effectiveness of machine learning models can be increased by analyzing how the features that build the model contribute to the detection of malware. The model can be made robust by getting insight into how features contribute to each sample that is fed to a trained model. In this paper, the boosting machine learning model based on LightGBM is enhanced with Shapely value to detect the contribution of the top nine features for classification such as true positive or true negative and for misclassification such as false positive or false negative. This insight in the model can be used for effective and robust malware detection and to avoid wrong detections such as false positive and false negative. The comparison of the top features and their contribution in shapely value for each category of the sample gives insight and inductive learning into the model to know the reasons for misclassification. Inductive learning can be transformed into rules. The prediction by the trained model can be re-evaluated with such inductive learning and rules to ensure effective and robust prediction and avoid misclassification. The performance of models gives 98.48 at maximum and 97.45 at a minimum by 10 fold cross-validation.

Keywords—Artificial intelligence; machine learning; malware detection; shapely value; decision plot; waterfall plot

I. INTRODUCTION

At the current time, the malware is generated in large numbers. Open Threat Exchange [1] is a platform for the exchange of information related to computer security. The reason for the high volume generation of malware is both from the generation side, and end-use of it. Malware authors use tools such as polymorphic and metamorphic engines. Metamorphic engines can generate malware with minor modification of code. It uses techniques such as register reassignment, NOP instruction insertion, code transposition, the substitution of machine-level opcode/instructions, dead code insertion, and combinations of these techniques. Polymorphic engines can generate malware with encryption, prepend data, append data, and combinations of these techniques. The generated malware exhibits the same behavior as old malware. However, this generated malware can evade detection by antivirus software based on the signature. The detection engine of many antiviruses is based on the signature. Hence, databases of signatures need a constant update for upcoming malware. On the use side of malware, the number of software products has

increased over time. Ten top software products with vulnerabilities are listed in Table I [2]. Software products with vulnerabilities from the top ten vendors are listed in Table II [3]. These vulnerabilities are exploited for an attack using existing or new malware. The software products are not limited to but include Operating Systems (OS), Driver for hardware devices, software applications, etc. The more a software product is used and popular, the more attacks it may have. Hence, hackers need more malware to attack the vulnerabilities. The vulnerabilities in hardware, OS, application, firewalls, anti-virus products, etc. may be by accident. The author [4] identifies three phases of the life cycle of vulnerabilities. In the first phase, a product is released in the market. The second phase starts when a vulnerability is found in the software product. In the third phase, the vulnerability has to be fixed by the developer and released for the user of the software. The vulnerabilities can be systematically discovered with needful tools. Knowing vulnerabilities is not enough, the vulnerabilities have to be proven by exploits, and attack software (malware). Machine learning and deep learning methods are used for malware detection and classification in research work these days.

TABLE I. TOP SOFTWARE VENDORS WITH VULNERABILITIES

SL No	Vendor Name	Number of Products	Number of Vulnerabilities	#Vulnerabilities/#Products
1	Microsoft	655	8178	12
2	Oracle	938	8043	9
3	Google	124	6571	53
4	Debian	106	5697	54
5	Apple	139	5380	39
6	IBM	1314	5334	4
7	Cisco	5592	4137	1
8	Redhat	407	3984	10
9	Canonical	49	3075	63
10	Linux	23	2751	120

TABLE II. TOP OPERATING SYSTEMS WITH VULNERABILITIES

Sl. No.	Product Name	Vendor Name	Product Type	Number of Vulnerabilities
1	Debian Linux	Debian	OS	5572
2	Android	Google	OS	3875
3	Ubuntu Linux	Canonical	OS	3036
4	Mac Os X	Apple	OS	2911
5	Linux Kernel	Linux	OS	2722
6	Fedora	Fedoraproject	OS	2538
7	iPhone OS	Apple	OS	2522
8	Windows 10	Microsoft	OS	2459
9	Windows Server 2016	Microsoft	OS	2233
10	Windows 7	Microsoft	OS	1954

The objective of this paper is to further improve the effectiveness of the machine learning (ML) model based on boosting algorithms such as LightGBM by overcoming the wrong prediction, misclassification the ML model may have. Good ML models are made general with feature engineering and learning from a large dataset, to detect unknown malware. There are many algorithms for ML models and many feature engineering techniques to make the models effective that resulting in increasing the accuracy of models. Misclassification in the machine learning model is wrong identification. For malware, the ML model may not identify them and they are termed as a false negative. A false negative detection can be very dangerous for any organization. As the malware is not detected, it will be able to meet the objective of the attacker despite all the security solutions applied. ML model may also declare benign software as malware. Such occurrences are termed false positives. A false positive detection causes issues such as panic among users of the software, inconveniences, non-use of software until a confirmed source declares the software as benign. All machine learning models have misclassification without exception.

Machine learning models to detect malware are many and they also use feature importance as part of an algorithm to identify top features. There are other methods for feature importance using feature engineering such as Principal Component Analysis (PCA), Redundant Feature Removal (RFR), and Haar Wavelet Transform (HWT) [6] and Leave One Feature Out importance (LOFO) method [7].

In this paper, a novel method is proposed to identify the change in top features that contribute to the misdetection of malware or future input sample that may be malware or benign software to a trained ML model. In addition, to identify the amount of contribution the top features are having for misclassification of a future sample in consideration as input to the ML model. Shapely values and visualization techniques are used to achieve these objectives. Shapely values are from classic game theory. Shapely values are used to find feature importance in an ML model. Lundberg et al. [5] have used Shapley value for explainable artificial intelligence. Hence, Shapley values can identify the top features in an ML model. The top features in an ML model based on LightGBM have

shapely values associated with them. These top features along with their contribution to Shapley values are visualized using decision plot, waterfall plot, and force plots. Further, this work proposes to identify the false positive and false negative from the test dataset part. Further, the work also associates visualization with change in top features and amount of contribution of top features. Having identified top features and the amount of contribution of the top features for misclassification, this work proposes the use of inductive learning techniques to overcome the misclassification of future samples. The present work aims to improve the effectiveness of the ML model based on the LightGBM model. It can be used for zero-day malware detection as well.

The gaps that this work addresses are highlighted as follows.

- These feature importance from algorithms and feature engineering methods cannot associate the top features for a new sample used for prediction by a trained machine learning model.
- They cannot determine the amount of contribution of a feature for a sample used for prediction by a trained machine learning model. Hence, they cannot associate the visualizations with the amount of contribution of a feature for a new sample to be predicted by the machine learning model.
- There remains always a doubt if the new sample under test is part of high accuracy as published for the model or part of misclassification as false negative or false positive.
- The inductive method proposed in this work improves the probability of prediction to a higher level.
- A novel approach as proposed in this work is not available in the literature survey. Hence, this paper opens new dimensions for increasing the probability of effective detection of a new sample by a trained model.

Specific contributions in this study are:

- Use of Shapley values and visualization for identification of top features for false negative (FN), false positive (FP), true positive (TP), and true negative (TN) categories of samples for LightGBM machine learning models.
- Amount of contribution by top features for each predicted category in Shapley values are identified. So that the comparison for inductive learning is effective.
- Comparison of the features and amount of contributions of features for samples with the test dataset part that may be FP, FN, TP, and TN. Using the comparison to identify the top features and their contribution for misclassified FP and FN samples.
- Use of LightGBM, boosting algorithms, for effective prediction of a future sample that may be malware or benign software. The proposed inductive method will avoid misclassification and improve the effectiveness of the ML model.

This paper is organized with a literature survey in Section II, followed by the methodology of malware detection and the use of shapely values for visualization in Section III. The Dataset, experimental setup, and results are outlined in Section IV. The paper concludes in Section V with a conclusion and an Appendix in Section VI.

II. LITERATURE SURVEY

Malware is like any software product. It has to be distinguished from a benign software product. The methods available to distinguish and detect the malware are broadly categorized into static analysis, dynamic analysis, and hybrid analysis.

A. Static Analysis

In static analysis, the malware is not executed. Features for machine learning are extracted from the software without running the software, the sample under consideration. It has the advantage that the sample cannot infect the system used for extraction of features. All the software, malware, shared libraries required, and dynamic link libraries (DLL) have a header. For windows, the header of the executable is termed Portable Executable (PE) header. The features from the PE header of windows executables can be extracted as explained in [6][8]. In addition, features can be extracted using properties of the executable file as an object and are termed file-related features. File related features are not limited to but include a histogram of bytes in executable, the entropy of complete file entropy of various parts of files, strings embedded in the executable, N-grams [9] from byte code, N-grams from assembly code, N-grams from API calls, images of hex bytecode of a file [10][11], images of hex bytecode of different part of a file, etc. Many machine learning models and deep learning models use features with different combinations derived from static analysis[12]. However, malware authors use methods such as obfuscation [13], encryption of various types to evade feature extraction methods. The obfuscation and encryption methods are many and may be categorized into standard and non-standard (private). These shortcomings of static analysis may be overcome by dynamic analysis [14].

Authors in [15] convert the sample file to images and extract features using the trained CNN model. The extracted features are plotted using t-Distributed Stochastic Neighbor to identify the cluster of malware. Subsequently, they make N-grams with n values 1 to 5 using the API call sequence for six types of malware actions. The malware actions are creating or modifying files, hooking on to system services, getting information for loading the DLL, etc. The N-grams are used with eight types of distance measurement to make a similarity matrix using four types of kernel functions with the Support Vector Method (SVM). Distance measurements used in this work are Cosine, Bray-Curtis, Canberra, Manhattan, Chebyshev, Euclidean, Hamming distance, and Correlation for feature extraction. This technique may handle malware with a known packing method, as they can be unpacked to process and get features but will have a deficiency in handling packed malware with unknown packing methods.

Yousefi-Azar et al in [15] extract static features of a sample, malware, or benign software, using term frequency based on

natural language processing. Extracted features are used with the deep learning model and Extreme Learning Machine (ETM) for malware detection. Backpropagation results in large feature space which increases computation complexity. The authors multiply term frequency with a random projection matrix to reduce the computation complexity. Balanced android dataset Drebin and Dexshare and windows executables from 2016 are used as a dataset. Windows executables from 2017 are tested as zero-day malware to achieve an accuracy of 95.5%.

The authors [16] collect malware samples that are used for attacks in financial institutions in Brazil, affecting cyber users for over 6 years. They use static analysis to extract features from PE header of collected samples and use Multilayer Layer Perceptron, K-nearest neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM) classifiers to detect malware. Further, they identify the family of malware using the t-Distributed Stochastic Neighbor Embedding (SNE) method. Concept drift of ML model is detected using Drift Detection Method (DDM) and Early Drift Detection Method (EDDM) to detect drift in the malware samples over time. The authors visualize and relate the new malware families coming over time using confirmation and warning indicated by the drift methods. They conclude that a warning indication by drift methods implies a degradation of ML models and a confirmation indication by drift method implies that the ML model needs to be updated.

B. Dynamic Analysis

In dynamic analysis, the malware is executed in a protected environment, and the behaviors, actions of malware are observed. In a normal environment, the sample will infect the system and will affect the future normal use of the system. Hence, a protected environment is used to avoid infection of the system conducting the malware test. The actions and behaviors of malware are not limited to but include adding, deleting, and modifying related changes in the file name, registry, processes, communication in the network, system configuration, etc. Features are derived with these changes and used in machine learning models with various algorithms. The dynamic analysis method is very expensive in terms of time to execute malware, computing resources, and trained manpower required. Besides, the malware authors employ techniques to avoid malware detection. One of the techniques employed by the malware author is to detect the virtual environment required for running the malware. If the virtual environment is detected, they switch off the behavior of malware and act as benign software. Another technique used by malware authors is to connect to the command and control center owned by them and download the malware at a later time to take control of the target machine. If the network is not available in virtual environment, the sample acts as benign software. Hence, trained persons are required to note this behavior of malware. The hybrid analysis is used to overcome these shortcomings of dynamic analysis.

Robert et al. [17] use a large dataset of malware with a Malheur tool to know the behavior of samples. Malheur tool executes the samples and generates a report. Useful information such as DLLs imported, API used as the callback are extracted from the report to understand the actions, behavior of malware with help of domain experts. Domain experts make rules and rules are externalized to the malware detection

module. Authors believe malware will exhibit its behavior as per framed rule and that can be detected. However, new types of malware may not exhibit behavior as per rules framed, because that malware was not part of the dataset used. Hence, this unknown malware will not be detected.

Binayak et al. [18] create a knowledge database of In-memory processes based on the use of Dynamic Link Library (DLL) sequences using TF-IDF (Term Frequency-Inverse Document Frequency) and multinomial logistic regression based learning approach. The suspected process from malware uses a different DLL than of system DLL. This knowledge database is compared with DLL sequences used by In-memory processes to identify suspected, unwanted processes and malware.

C. Hybrid Analysis

Hybrid analysis combines static and dynamic analysis to overcome their shortcoming. Lifan Xu et al. [19] extract both static and dynamic features from android malware dataset and represent the features as vector. Advance features are derived using deep learning, a Deep Neural Network (DNN) using both the original static and dynamic feature vector sets. The advanced and original features are concatenated as new vectors as input to the DNN that modifies with multiple different kernel to detect malware. The combined hybrid analysis has shortcomings as in dynamic analysis or static analysis.

Sethi et al. [20] use feature from both static analysis and dynamic analysis on PHP, pdf, exe files. For dynamic analysis, the authors use a Cuckoo sandbox. Cuckoo sandbox is a virtual environment to run executable. It gives an analysis report of actions and behavior of the file executed. J48, SMO, and Random Forest machine learning algorithms are applied in the WEKA tool with the combined feature extracted using static and dynamic analysis. They achieved 100% accuracy with J48.

The literature survey gives different methods of improving the accuracy and other performance parameters of the machine learning model by feature engineering for malware detection. However, they do not give insight into the top features and contribution of each feature for a new sample by a trained machine learning model. Hence, there is a gap in research that can give insight into the top features and their contribution in the prediction of an unseen sample by machine learning model. This work is an effort to fill the gap.

III. METHODOLOGY

A. Shapley Value and Feature Importance

The machine learning model should be both interpretable and accurate. Interpretation of ML model based on decision tree may be based on decision path, heuristic value to features, and model-agnostic. In this work, Shapley value is used for making the ML model interpretable. A local explanation is assigning a numeric measure, credit, to each input feature that constitutes a machine learning model based on a decision tree. These local explanations are combined to represent a global structure that represents an ML model based on a decision tree or an ensemble of decision trees. The ensemble of decision trees may be based on a bagging algorithm such as Random Forest or boosting algorithm such as LightGBM. The global explanation

of the ML model continues to retain the local faithfulness as in local explanation. Shapely values from game theory satisfy simultaneously local accuracy, consistency, and missingness three properties required for credit score to a feature in an ML model. The credit score, Shapley values, are computed by one feature at a time into the output function of the model with some condition as in Eq. (1). Lundberg et al. [5] follow the causal do-notation formulation. It justifies use of the Shapley additive explanation (SHAP) interaction values as a richer type of local explanation and feature perturbation formulation.

$$f_x(S) = E[f(X)|do(X_s = x_s)] \quad (1)$$

S = Set of features to condition on

X = A random variable from M input features of model

x = input vector for the current prediction for the model

Lundberg et al. [5] give TreeExplainer, an explanation method for ML models based on tree, that enables optimal local explanations based on shapley values from classic game theory. Classic Shapley values are ways to measure feature importance. It is optimal and maintains natural properties from cooperative game theory. Exact computation of these values is NP-hard problem. Hence, they have approximate computation. Authors have developed an algorithm for decision tree categories of algorithms that computes local explanations with theoretical guarantees of local accuracy and consistency in polynomial time with Shapley values. Local explanations are also used to capture feature interactions in a theoretically grounded way. S is the set of features in Eq. (1) to condition on and refers to features of a specific tree in the ensemble of trees in the boosting LightGBM machine learning model. We can find the SHAP value for each feature, x in Eq. (1), in a tree using the TreeExplainer and add for all the features, X in Eq. (1), in the tree under consideration to match with the tree. This can be applied to all the trees in the model one by one. Finally, we find the contribution of a feature for the ensemble of trees by the TreeExplainer. By knowing the contribution of all features in an ML model, it provides valuable insight into top features for each prediction.

B. Malware Detection Model

All samples in the dataset are from windows executable. The features are derived from the PE header of the samples and as properties of a file. Each window executable contains a PE header that is explained in [6] [8] [21]. The PE header can be extracted using a python program using "Library for Instrumenting Executable Files" (LIEF) a library in python. The extracted features are listed in detail in Appendix A. PE header consists of DOS header, file header, NT header, section header, optional header, and many directories such as Import directories, Resource directory, Export directory, and Exception directory. Import directories list Dynamic Link Libraries (DLL) loaded by the executable and Application Program Interfaces (APIs) used by executables. Resource directory lists the information required by executable such as icons, bitmaps, strings, menus, dialogs, configuration files, version information, etc. Exception directory lists exception handling information. Features extracted are listed in Appendix A. Some of the features are described here. File header of PE header gives features such as timestamp, vsize, has_debug, has_relocations,

has_signature, has_tls, has_symbol, imports, Machine1-Machine10 listed in Appendix A. Machine representing, type of processor required, in the file header part of PE header is hashed and put into one of ten bins and named as Machine1 - Machine10. Features that are hashed and put in several bins are named like this. Section header and optional header give section name, section size, section characteristics, and start and end byte contents of each section. The section name is a string. It is hashed and put into 1 of 50 bins. This gives us feature entry_name1 - entry_name50 listed in Appendix A. Section size, section virtual size, and section characteristics values are hashed and put into 1 of 50 bins. These operations give us features Sec_size_1 - sec_size_50, sec_vsize1 - sec_vsize50, sec_char1 - sec_char50 listed in Appendix A. Entropy of content of each section in the sample is hashed and put in to 1 of 50 bins. This gives us features sec_entropy_1 - sec_entropy_50 listed in Appendix A.

DLL in an import directory and the name of an API in the DLL are concatenated to make a string. The string is hashed and put into one of 1280 bins. This gives us the feature Imp1-Imp1280 listed in Appendix A. Function name in the export directory is hashes and out into one of 128 bins. This gives us feature exp1-exp128 listed in Appendix A. File-related information used to derive features are histogram of bytes, strings, and entropy of hex values in each sample. The byte value in the sample can be 0-255. A histogram is count of value of the byte in each sample. The count of the value of a byte is put into the respective bin H1-H256 to represent the feature listed in Appendix A. Strings in a sample give very important, insightful information used by malware. Strings reveal created and modified filenames and registry related information. Strings may also reveal IP addresses used by malware authors for communication, command and control center URLs, signature of malware authors and groups. All strings of size five-character or more are extracted, hashed, and put in one of 104 bins. This gives us features Str1 - Str104 listed in Appendix A. The encryption and packing methods increase the entropy, disorder of bytes in samples. Entropy is computed as the method described by [8]. In this method, a block size of 2048 bytes is extracted and counts of bytes are put in 16x16 bins. These operations of making a block of 2048 bytes with windows of 1024 bytes and putting in 16x16 bins are repeated for the entire content of a sample. This gives us features Ben1-Ben256 listed in Appendix A. Both the PE header and file-related information give 2351 features. Dataset consists of malware and benign samples and belongs to January 2017 time period.

Gradient Boosting Decision Tree (GBDT) LightGBM ML algorithms are selected for experiments in this work. The ML algorithm is selected for the following advantages.

- Feature importance of the ML model can be extracted after training of the model.
- Faster training and prediction
- Ease of computation

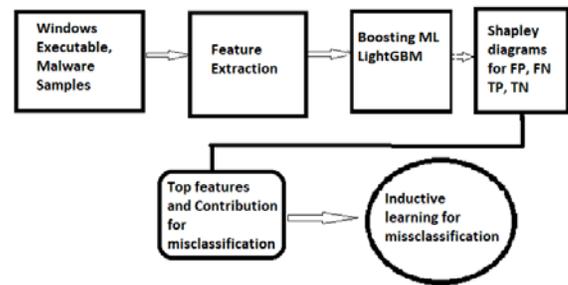


Fig. 1. System Block Diagram for Top Feature and their Contribution from Shapley Diagrams for Misclassification and Inductive Learning.

The system block diagram for this work is shown in Fig. 1. LightGBM boosting machine learning algorithm in the sklearn library is used to train the ML model. A trained model can predict the samples in the test dataset and correct detection of samples in true positive (TP), malware, and true negative (TN) benign software categories. Misclassified samples such as false positive (FP), benign software detected as malware, and false negative (FN), malware detected as benign software categories can also be identified. Nine to twenty five top features among the 2351 features can be identified for samples in TP, TN, FP, FN categories using diagrams such as waterfall plots, decision plots, and force plots. These diagrams show the amount of contribution by each top feature in Shapley values. Shapley values give a local explanation of top features with global structure as per ML model prediction for a sample. Changes in the top few features and their contribution to Shapley value for TP, TN, FP, and FN is compared. The comparison identifies a change in top features and their contribution for FP, FN also. This insight can be used as inductive learning to identify other samples which may have been misclassified and for future unknown samples without labels. Having found the misclassified samples by trained ML classifier, correct classification or malware detection can be performed. This leads to an increase in the performance of the trained ML model. One has to be very careful in this comparison and inductive learning with an unknown sample that does not have a label. Top features and their contribution in Shapley value for the unknown sample should match top features and their contribution in Shapley value for with known TP and TN samples also.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Dataset

The dataset in the proposed system is derived from [21]. It has Malware data from December-2006 to December 2017. The dataset from December 2006 to December 2016 contains only the malware and no benign entries and the reason for exclusion. Fig. 2 shows the exclusions, filter and pre-process used on the dataset to get the sub dataset used in this experiment. Dataset part from January 2017 is used in this proposed system. The unidentified entries are without labels in the dataset and are excluded for malware detection and analysis. The unidentified entries in the dataset may be malware or cleanware. The dataset consists of 32761 malware and 17186 benign software that appeared in January-2017. The details of the derived dataset are in Table III.

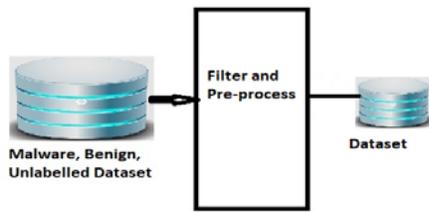


Fig. 2. Filter and Process the Database for the Experiments.

TABLE III. DATASET USED IN THIS WORK

SL No.	Samples	label	appeared
1	28606	Unidentified	2017-01
2	17180	Benign	2017-01
3	32761	Malware	2017-01

Each entry in the dataset has 2351 features. These features are from PE headers, sections of windows executable, systems APIs used in the executable, exported API from the executable, and file related properties. File related properties include Histogram of the complete executable in 256 bins, Byte entropy of executable file hashed into 256 bins and strings in 104 bins. The executable here means both the malware and cleanware. These features are defined in Appendix A and are used in the various diagrams in this paper. These features' names help identify exact features that are contributing to the detection of malware and the amount of contribution in the detection of malware or cleanware.

B. Experimental Setup

Intel(R) Core(TM) i5-7200U CPU @ 2.50 GHz, 2701 MHz, 2 Core(s), and 4 Logical Processor(s) with 8 GB Ram is used as computing resource in this work.

C. Malware Detection with LightGBM

Dataset is divided into a training set and testing set in the ratio of 70% and 30%. The model is trained with the training set and tested with the testing set. The results of this are in row 1 of Table IV. It has performance data for Accuracy, Precision, Recall, F1-score, and confusion matrix parameters in terms of false negative (FN), false positive (FP), true positive (TP), and true negative (TN).

30% of the dataset is separated for the testing of the LightGBM model. The samples in the test dataset are identified in false negative (FN), false positive (FP), true positive (TP), and true negative (TN) categories. It is interesting to explore how the top features and other features contribute to FP, FN, TP, and TN samples by the LightGBM algorithm. Waterfall plot, decision plot, and force plot are drawn with Shapley values.

TABLE IV. PERFORMANCE OF LIGHTGBM MODEL WITH ZERO-DAY MALWARE

	Accur acy	TP	F P	F N	TN	precis ion	rec all	f1- score	supp ort
D1- Test	98.483	552 0	15 0	10 0	107 11	0.99	0.9 9	0.99	1081 1
D1	99.389	169 98	18 2	12 3	326 38	0.99	1	1	3276 1

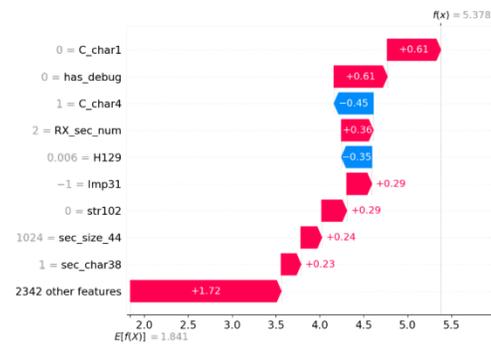


Fig. 3. Waterfall Plot of True Positive Sample in a Dataset with Shapley Values for each Feature.

Fig. 3 shows the waterfall plot for a true positive in the dataset. The Shapley value sum features to 5.378. The waterfall plot adds the contribution of each feature and also shows the top features with their contribution leading to the decision. Although the total contribution of 2342 features; lowest bar, in figure is significant compared to any top feature. In additional analysis, the feature importance of LightGBM showed only 588 features contributed to the model. Other 1763 features do not have any contribution to malware detection. Hence, many features in 2342 features have zero contribution. Kumar et al. [22] identified that 276 features among 2351 only contributed to prediction in the XGBoost model with 600k samples of training dataset from [21]. The remaining 2075 features have zero contribution to the model.

These figures help us to identify top features contributing to decision at leaf node with LightGBM algorithm.

Fig. 4 displays the decision plot for a true positive entry in the dataset and shows how the top features contribute to make the decision. The decision plot adds the contribution of each feature and draws the line that takes it to make a decision. It has the same TOP features as in Fig. 3. The waterfall plot adds the contribution of all the features but does not show the graph that leads to a decision as in the decision plot.

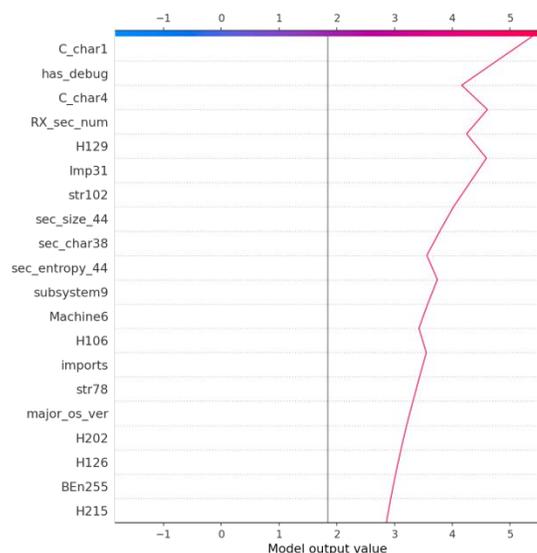


Fig. 4. Decision Plot of True Positive Sample in Dataset with Shapley Values for each Feature.

The decision plot can be for more number of samples. Fig. 5 shows the decision plot for the first ten samples in the dataset. The value that shows negative from zero in blue color are benign and the values that are on the right side of two, the purple color vertical line, in the graph are malware. The seven samples, pink color, with the specific features as shown are the malware and three samples are benign. The objective of these figures is to display how the features are contributing to decision with use of the LightGBM model. The label of samples is verified with the prediction of each sample with the LightGBM GBDT algorithm for all the 10 samples. It matches as given in the decision plot.

Fig. 6 shows the force plot for a true positive sample in the dataset in the Shapley values. The force plot shows how each of the feature is contributing in the positive direction from the left side with red color and other features that contribute negatively to scale value down and finally the value set near 5.38 for Shapley values. The top three features that contribute to the decision are named. The meaning of these features can be referred at Appendix A. The force plot cannot display the name of many features as in the decision plot. The top three features are the same as in Fig. 3 and Fig. 4.

Fig. 7 shows the force plot for the first 10 samples of the dataset in Shapley values. This figure is like rotating Fig. 6 clockwise and stacking the ten force plot of the figure in the x-axis. The count of the sample is seen at the top with numbers 0, 1, 9. The Y-axis displays the Shapley value for samples. The Shapley values for sample 2 are different from another sample. This change in Shapley value for each sample in the dataset is visible. It is possible to change the parameters in x-axis and y-axis from a drop down menu and analyze the top features for a sample. Amount of contribution top features make to the decision using the LightGBM algorithm can be observed.

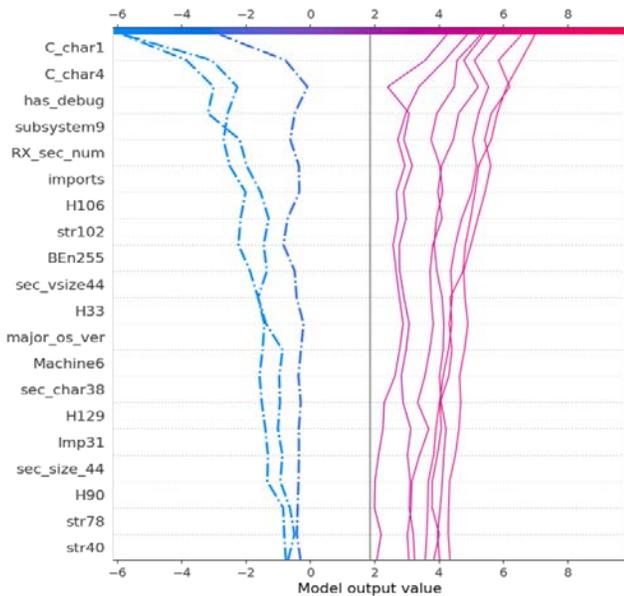


Fig. 5. Decision Plot of First 10 Samples in Dataset with Shapley Values for each Feature.

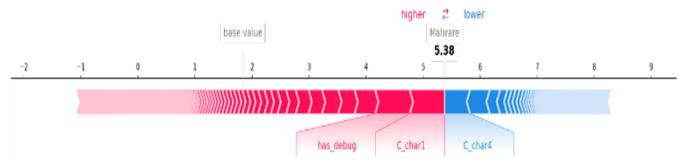


Fig. 6. Force Plot of True Positive in Dataset with Shapley Values for each Feature.

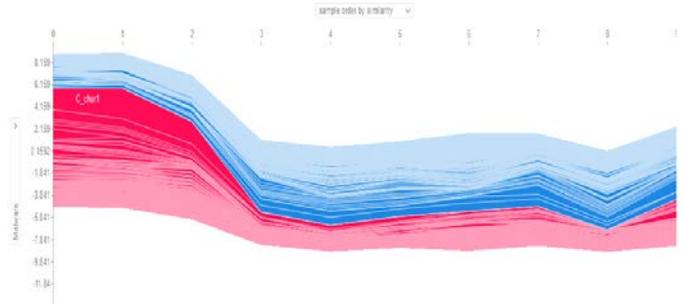


Fig. 7. Force Plot of First 10 Samples in Dataset with Shapley Value for each Feature.

Fig. 8 displays a waterfall plot for a false positive sample from the test dataset part of the dataset. The advantages of waterfall plots are:

- Top 9 features contributing to predicting the sample as a false positive.
- In the Shapley scale, it starts at 2.0 and adds up to 3.589 with the top 9 features contributing in both positive and negative directions.
- The contribution of the remaining 2342 features for the sample is +0.2, much less than the top five features.

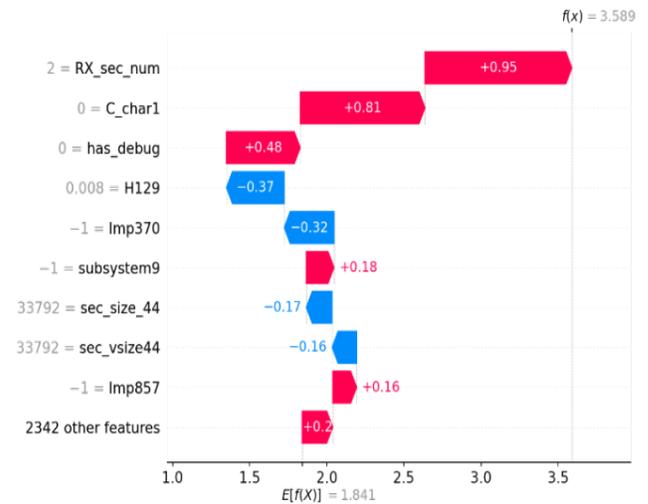


Fig. 8. Waterfall Plot of a False Positive Sample in Dataset in Shapley Value for each Feature.

Fig. 9 reveals the force plot for a false positive sample in the test dataset from dataset. The top features are RX_sec_num, C_char1, and has_debug from the PE header. The contribution of RX_sec_num is the highest among all the features. The top features for false positive in Fig. 9 are very different from the top features of malware (true positive) in Fig. 3. In addition, the final Shapley value for the sample is down to -0.09 in Fig. 9 compared to 5.38 in Fig. 3 for malware. The start point is very low at less than -6 in Fig. 9 compared to the start point at -1 in Fig. 3.

Fig. 10 presents a waterfall plot for a False Negative sample in the test dataset from the dataset in the Shapley value. All the advantages as explained for a false positive sample can be observed. In addition, features that are making the sample false positive and false negative can be compared. There is no contribution of the RX_sec_num, H129 feature in the false negative sample as in the false positive sample. The contribution of feature C_char4 is there in false negative but not in false positive.

Fig. 11 shows the force plot for a False Negative (FN) sample in the test dataset in Shapley values.

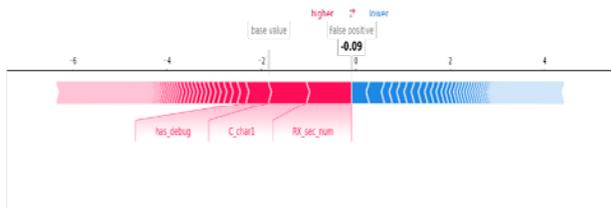


Fig. 9. Force Plot of a False Positive Sample in Dataset in Shapley Value for each Feature.

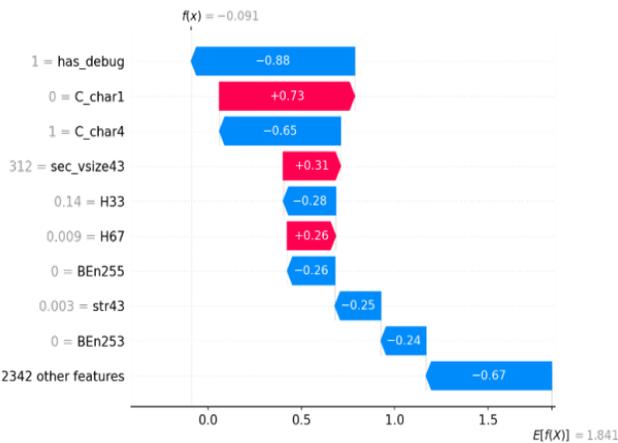


Fig. 10. Waterfall Plot of a False Negative Sample in Dataset in Shapley Value for each Feature.

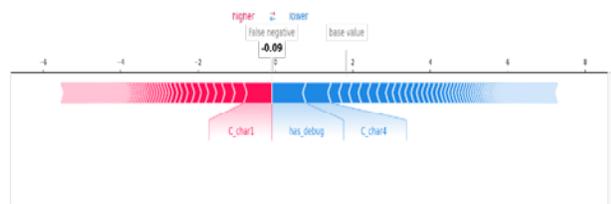


Fig. 11. Force Plot of a False Negative Sample in Dataset in Shapley Value for each Feature.

Fig. 12 shows the waterfall plot for the True negative sample in test dataset in Shapley value. Fig. 13 gives the force plot for the True negative (TN) sample in the test dataset in Shapley value.

The top features of waterfall plots in Shapley value from Fig. 3, Fig. 8, Fig. 10, Fig. 12 for in FP, FN, TP, and TN samples respectively in test dataset of the dataset are compared in Table V. Top features are listed in the features column. For a sample in each category in false positive, false negative, true positive and true negative, it identifies the presence of a feature as “Y” and no presence as “N”. Further, it identifies the topmost feature, with the value among the top feature with a “T” in each category. The probability value contributed by each feature is identified in respective columns. This table helps to conclude that there is disjoint set of features for each category samples in FP, FN, TP, and TN. The topmost feature for the FN sample is has_debug and is present in FP and TP. The topmost feature for FP is Rx_sec_num and contributes very low value in other categories of samples.

The contribution of the remaining 2342 features is lowered significantly for FP and FN. For TP the value is +ve .42, for TN the value is negative -.03.

These comparisons can identify the misclassified FP and FN samples and improve the efficiency of the ML model by correct classification for an unknown sample. Few insightful rules that can be formed are as follows:

- The malware sample with a high contribution of Imp321, H33, C_char1, and str43 may be a FP sample.
- The Malware sample with the highest contribution by Rx_sec_num among all the features will be a FN sample.

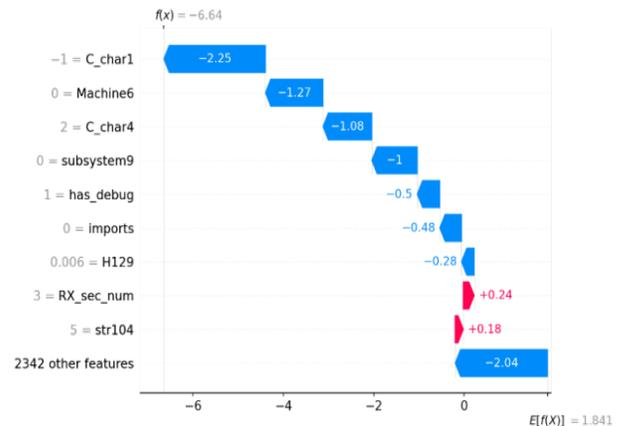


Fig. 12. Waterfall Plot of a True Negative Sample in dataset in Shapley Value for each Feature.

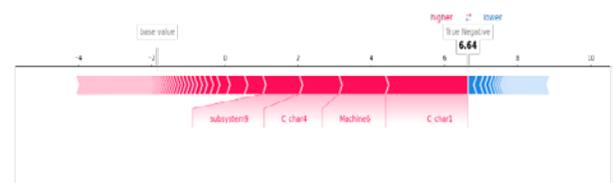


Fig. 13. Force Plot of a True Negative Sample in the Dataset in Shapley Value for each Feature.

TABLE V. FEATURES AND THEIR CONTRIBUTION IN FALSE POSITIVE, FALSE NEGATIVE, TRUE POSITIVE, TRUE NEGATIVE PREDICTION BY LIGHTGBM MODEL

Sl no	Features	False Positive	False Negative	True Positive	True Negative
1	C_char1	Y,+0.81	Y, 0.73	Y, 0.61	Y, -2.25
2	C_char4	N	Y, -0.65	Y, -0.45	Y, - 1.08
3	Machine6	N	N	N	Y, -1.27
4	Subsystem9	0.18	N	N	Y, -1.0
5	Rx_sec_num	Y, +0.95	N	Y, 0.36	0.24
6	has_debug	Y, +0.48	Y, -0.88	Y, 0.61	-0.5
7	sec_size43	Sec_size44 = -0.17	Y, 0.31	0.24, sec_char38	N
8	H129	Y, -0.37	Y, H33 = -0.28 H67=0.26	Y, -0.35	-0.28
9	Imp370	Y, -0.32 imp857=0.16	N	Imp31=0.29	N
10	sec_char38	N	N	0.23	N
11	Str43	N	-0.25	N	str104=0.18
12	imports	N	N	N	-0.48
13	Ben255	N	-0.26, BEn253=-0.24	N	N
14	Other 2342 features	0.2	Y,-0.67	Y, 1.72	Y, -2.04

D. k-Fold Cross-Validation

Cross-validation with k=10 is performed for less biased and less optimistic accuracy value for the LightGBM model. The test dataset is used for this 10-fold cross-validation test. The results of cross-validation are tabulated in Table VI.

TABLE VI. TEN K FOLD CROSS-VALIDATION FOR THE TEST DATASET

Sl. no	Accuracy
1	0.97938144
2	0.9836165
3	0.97936893
4	0.9848301
5	0.97451456
6	0.97815534
7	0.97512136
8	0.97936893
9	0.9836165
10	0.97815534

E. Comparison with other Malware Detection Works

This work is compared with other malware detection works in Table VII. This work achieves higher accuracy with datasets compared to Yousefi-Azar et al. and comparable accuracy with Venkatraman et al. and Alazab et al. Jung et al [24] take 333 malware files with .swf extension into the test dataset from 2007-2015 for zero-day malware to get 51–100 % accuracy. Alazab et al. [25] get marginal higher accuracy of 98.6 compared to 98.49%. Shafiq et al. have a small size dataset and give the model performance at 99.2 Area under curve (AUC) that cannot be compared with accuracy.

They use more than three times malware compared to benign software for training and testing. They do not define ways to determine unknown malware. [6] Use only one tenth of benign software compared to malware. This highly unbalanced dataset lowers the probability of false positives. They consider zero-day malware as one which does not match known signature or unknown malware.

TABLE VII. COMPARISON WITH OTHER ZERO-DAY MALWARE WORKS

Paper	Method	Sample/Dataset	Result/Accuracy
This work	Boosting algorithms: LightGBM	Dataset Details in Table III	98.49
Yousefi-Azar et al. [15]	NLP and the term frequency tf-simhasing: term frequency of sample multiple with rand projection matrix	Android:Drebin, DexShare Windows PE files: Training:11983 Malware, 8912 Benign (2016) Testing: 12127 Malware, 11983	97.33
Venkatraman et al. [23]	Malware files to image as input to pre-trained CNN to get features, Apply SVM with SMO-Normalized Polynomial	52k samples	98.6%
Jung et al. [24]	API call sequence features Use Deep Feed-forward NN, RNN	Malicious .swf files 333 Benign .swf files 333	51% to 100%
Alazab et al. [25]	NB, kNN, 4 kernels with SMO. SMO–PolyKernel, SMO –Puk, SMO-Normalized, and SMO- RBF Backpropagation J48 and Neural Networks Algorithm	66703 samples with 51223 Malware and 15480 Benign software	98.6
Shafiq et al. [6]	Ripper, Ibk and SVM-SMO classifier	1447 Benign software 8892 + 5586 = 14478 malware	99.2% g AUC

V. CONCLUSION

In this work, a boosting machine model based on LightGBM is enhanced using Shapely value to build an effective and robust machine learning model. Features derived by static analysis of malware and benign samples in the dataset are used to build the LightGBM boosting machine learning model. Datasets from Jan 2017 for malware is used for training and prediction. Waterfall plots, Decision plots and Force plots based on Shapely value helped identify the top few features. The Waterfall plots demonstrated a change in features and their contribution for a sample from different categories of samples as insight into the ML model. Table V compared the top features contributed to misclassified samples. The top feature for samples that is detected as false positive, false negative samples by trained models is analyzed and inductive learning rules are made. The inductive learning rules can be applied to unknown, unlabeled samples to avoid misclassification into FP and FN and to ensure correct detection. These top features and their contribution may be used to overcome the misclassification of malware. The cross-validation with the test dataset is 98.48 at maximum and 97.45 at minimum.

The work can be further extended to analyze change in features and to derive inductive learning rules for misclassification by other ML models for false positive and false negative cases to ensure correct prediction. The Shapely values for a feature may be mapped to the probability score of the ML model. This will help to correlate the Shapely value to probability value for a feature as local explanation and as a whole for a sample at global explanation (structure). Large datasets may be used to make a robust ML model and analyze reasons for misclassification for various families of malware such as ransomware, rootkit, Trojan horse, etc.

REFERENCES

- [1] "AlienVault - Open Threat Exchange." <https://otx.alienvault.com/> (accessed Dec. 29, 2021).
- [2] "Top 50 products having highest number of cve security vulnerabilities." <https://www.cvedetails.com/top-50-products.php> (accessed Dec. 29, 2021).
- [3] "Top 50 Vendors By Total Number Of 'Distinct' Vulnerabilities." <https://www.cvedetails.com/top-50-vendors.php> (accessed Dec. 29, 2021).
- [4] H. Pohl, "Zero-Day and Less-Than-Zero-Day Vulnerabilities and Exploits," 2008.
- [5] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020, doi: 10.1038/s42256-019-0138-9.
- [6] M. Z. Shafiq, S. M. Tabish, F. Mirza, and M. Farooq, "A Framework for Efficient Mining of Structural Information to Detect Zero-Day Malicious Portable Executables," no. October, 2015.
- [7] S. A. Roseline, S. Geetha, and S. Member, "High Performance Android Malware Detection System using Gradient Boosting based Static Feature Selection and Classifier Paradigm," pp. 1–25.
- [8] Stamp, Mark, Mamoun Alazab, and Andrii Shalaginov. *Malware Analysis Using Artificial Intelligence and Deep Learning*. Springer, 2021.
- [9] J. Saxe and K. Berlin, "Deep neural network based malware detection using two dimensional binary program features," 2015 10th International Conference on Malicious and Unwanted Software, MALWARE 2015, pp. 11–20, 2016, doi: 10.1109/MALWARE.2015.7413680.

- [10] S. A. Roseline, S. Geetha, S. Kadry, and Y. Nam, "Intelligent Vision-Based Malware Detection and Classification Using Deep Random Forest Paradigm," *IEEE Access*, vol. 8, pp. 206303–206324, 2020, doi: 10.1109/ACCESS.2020.3036491.
- [11] J. Hemalatha, S. A. Roseline, S. Geetha, S. Kadry, and R. Damaševičius, "An efficient densenet - based deep learning model for Malware detection," *MDPI Entropy*, vol. 23, no. 3, pp. 1–23, 2021, doi: 10.3390/e23030344.
- [12] S. K. J. Rizvi, W. Aslam, M. Shahzad, S. Saleem, and M. M. Fraz, "PROUD-MAL: static analysis-based progressive framework for deep unsupervised malware classification of windows portable executable," *Complex & Intelligent Systems*, Oct. 2021, doi: 10.1007/s40747-021-00560-1.
- [13] D. Mafaz, "Generic Packing Detection Using Several Complexity Analysis for Accurate Malware Detection," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 1, pp. 7–14, 2014, doi: 10.14569/ijacsa.2014.050102.
- [14] M. Tang and Q. Qian, "Dynamic API call sequence visualisation for malware classification," *IET Information Security*, vol. 13, no. 4, pp. 367–377, 2019, doi: 10.1049/iet-ifs.2018.5268.
- [15] M. Yousefi-Azar, L. G. C. Hamey, V. Varadharajan, and S. Chen, "Malytics: A malware detection scheme," *IEEE Access*, vol. 6, pp. 49418–49431, 2018, doi: 10.1109/ACCESS.2018.2864871.
- [16] F. Ceschin, F. Pinage, M. Castilho, D. Menotti, L. S. Oliveira, and A. Gregio, "The Need for Speed: An Analysis of Brazilian Malware Classifiers," *IEEE Security and Privacy*, vol. 16, no. 6, pp. 31–41, 2019, doi: 10.1109/MSEC.2018.2875369.
- [17] R. Gove, J. Saxe, S. Gold, A. Long, G. B. I. Labs, and Z. Piper, "SEEM: A scalable visualization for comparing multiple large sets of attributes for malware analysis," *ACM International Conference Proceeding Series*, vol. 10-Novembe, pp. 72–79, 2014, doi: 10.1145/2671491.2671496.
- [18] B. Panda and S. N. Tripathy, "Detection of Anomalous In-Memory Process based on DLL Sequence," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 185–194, 2020, doi: 10.14569/IJACSA.2020.0111025.
- [19] L. Xu, D. Zhang, N. Jayasena, and J. Cavazos, "HADM: Hybrid Analysis for Detection of Malware," *Lecture Notes in Networks and Systems*, vol. 16, pp. 702–724, 2018, doi: 10.1007/978-3-319-56991-8_51.
- [20] K. Sethi, B. K. Tripathy, S. K. Chaudhary, and P. Bera, "A Novel Malware Analysis for Malware Detection and Classification using Machine Learning Algorithms," *ACM International Conference Proceeding Series*, pp. 107–116, Oct. 2017, doi: 10.1145/3136825.3136883.
- [21] H. S. Anderson and P. Roth, "EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models," 2018.
- [22] Kumar, Rajesh; Geetha S, "Malware classification using XGboost-Gradient Boosted Decision Tree," *Advances in Science, Technology and Engineering Systems Journal*, Sep. 2020, doi: 10.25046/aj050566.
- [23] S. Venkatraman and M. Alazab, "Use of Data Visualisation for Zero-Day Malware Detection," *Security and Communication Networks*, vol. 2018, 2018, doi: 10.1155/2018/1728303.
- [24] W. Jung and S. Kim, "Poster: Deep Learning for Zero-day Flash Malware Detection," In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, pp. 2–3, 2015.
- [25] M. Alazab, S. Venkatraman, P. Watters, and M. Alazab, "Zero-day malware detection based on supervised learning algorithms of API call signatures," *Conferences in Research and Practice in Information Technology Series*, vol. 121, no. June 2014, pp. 171–182, 2010.

APPENDIX A

Bn1-Bn256: The entropy of executable for a window size of 2048 bytes is computed for the joint distribution of byte value and put into 16x16 bins. This is repeated for a step size of 1024 for the full file.

C_char1 - C_char10: Characteristics of the sample from PE file header to indicate if the file is DLL, Executable, systems file, etc. The value is hashed and put into one of the ten bins.

dll_c1 – dll_c10: The DLL characteristics value from the optional header of PE for the sample is hashed put into one of the ten bins.

entry_name1 - entry_name50: The name of each section in the PE header of the sample is hashed and put into one of the fifty bins.

exp1 – esp128: The exported APIs in the sample are hashed and put into one of the 128 bins.

Exports: Flag interpreted by LIEF, a python package, indicating the executable exports API in the data directory of PE header.

Imp1 – Imp1280: The DLL names and imported APIs in the DLL are hashed and put into one of the 1280 bins.

H1-H256: Byte count of hex value 0x00 to 0xFF of benign software, malware is put their respective bin. These counts are further normalized with the file size.\

has_debug: A flag in the characteristics field of file header in PE header of the sample, indicating debug information for the sample.

has_relocations: A flag in the characteristics field to indicate relocation sections, relocation directory.

has_resources: A flag in the characteristics field to indicate a resource section, a resource data directory.

has_signature: A flag in the characteristics field to indicate digital signature related information.

has_tls: A flag in the characteristics field to indicate tls section, tls data directory.

has_symbol: A flag in the characteristics field to indicate debug section with symbols.

Imports: Flag interpreted by LIEF, a python package, indicating the executable has imports of API from DLL.

Machine1-Machine10: It indicates hardware architecture 32/64 bit, processor for executable. The values are hashed and put in to one of the ten bins.

Magic1 – Magic10: Magic value from optional header of PE for the sample is hashed and put in to one of the ten bins.

num_of_sec_morethan0: Number of sections in section part of PE header which has content and size greater than 0 size.

Num_sec_noname: Number of sections in section part of PE header which are without a name. Generally, the name of a section is .text, .rdata, .data etc.

RX_sec_num: Number of sections in section part of PE header which has read and execute permission.

Sec_size_1 – sec_size_50: The size of each section in the PE header of the sample is hash and put in to one of the fifty bins.

sec_entropy_1 -- sec_entropy_50: The entropy of each section in the PE header of the sample is computed, hashed, and put into one of the fifty bins.

sec_vsize1 -- sec_vsize50: The memory size of each section in the PE header of the sample is hashed and put in to one of the fifty bins.

sec_char1 - sec_char50: The characteristics of each section in the PE header of the sample is hashed and put in to one of the fifty bins.

Size: Size of executable.

Str1-Str104: Five or more printable characters in the samples are hashed in to 104 bins. These strings include URLs starting with HTTP: HTTPS: registry keys starting with HKEYS, paths in systems such as c: /, file name, malware author's messages, etc.

Subsystem1 – Subsystem10: Subsystem value from optional header of PE header of the sample. The values are hashed and put in to one of the ten bins.

timestamp: Date, the timestamp of a sample.

Vsize: virtual size of executable in memory.

W_sec_num: Number of section in section part of PE header which has write permission.

Using a Rule-based Model to Detect Arabic Fake News Propagation during Covid-19

Fatimah L. Alotaibi, Muna M. Alhammad
Collage of Business Administration
King Saud University
Riyadh, Saudi Arabia

Abstract—Since the emergence of the Covid-19, both factual and false information about the new virus has been disseminated. Fake news harms societies and must be combated. This research aims to identify Arabic fake news tweets and classify them into six categories: entertainment, health, politics, religious, social, and sports. The study also aims to uncover patterns in the spread of Arabic fake news associated with the Covid-19 pandemic. The researchers created an Arabic dictionary and used text classification based on a rule-based system to detect and categorize fake news. A dataset consisting of 5 million tweets was analyzed. The developed model achieves an overall accuracy of 78.1% with 70% precision and 98% recall. The model detected more than 26006 fake news tweets. Interestingly we found an association between the number of fake news tweets and dates. The result demonstrates that as more information and knowledge about Covid-19 become available over time, people's awareness increase, while the number of fake news tweets decreases. The categorization of false news indicates that the social category was highest in all Arab countries except Palestine, Qatar, Yemen, and Algeria. Conversely, fake news related to the entertainment category was the weakest dissemination in most Arab countries.

Keywords—Fake news; Covid-19; text classification; rule-based system; trends

I. INTRODUCTION

Social media networks are increasingly being used as a source of information. Its low cost, easy access, and speed of information delivery encourage people to use it as a searching tool to obtain information. The widespread use of social media around the world provides a perfect setting for the dissemination of fake news. Fake news is described as low-quality news [1] or false news purposely broadcast to divert people away from true news and facts [2]. It is considered to be one of the most dangerous weapons capable of bringing harm to both society and people [3]. Its prevalence increases during crises as crises represent a perfect time for spreading fake news and rumors, especially when knowledge of a problem is limited and ambiguous.

Due to the breakout of the novel coronavirus "Covid-19" in 2020, the world faced health and economic concerns. The virus was first discovered in China in November of 2019 then formally declared a pandemic by WHO [4] on the 11th of March 2020, making it the worst global crisis of the 21st century. Many people are afraid and anxious because of the virus's secrecy and uncertainty. Therefore, both true and false news began to spread about this virus. Social network websites have become essential for obtaining news and information

about Covid-19 [5]. With curfew and lockdown being implemented in many countries, social network websites were full of endless discussion about the Covid-19, with much fake news being exchanged. Many individuals intentionally propagated social media with fake news to gain personal benefits, such as having a lot of likes and followers. In addition, some companies and organizations have benefited from the spread of false and misleading information on social media to promote and advertise products or services to increase their sales profit [6]. The propagation of fake news had a negative influence on public health throughout this crisis, increasing tension, rage, anxiety, panic, and depression [2]. It has been documented, for example, that some consumers in the United Kingdom and Australia experienced panic-buying while purchasing a specific product, such as toilet paper [7]. As a result, the negative impact of the propagation of fake news may lead to more severe problems within society as well as the Covid-19 problem.

As a response to control the spread of fake news during the pandemic, many countries and governments have taken the issue seriously and introduced new laws to prevent the spread of such information. For example, Twitter announced that it would remove any misleading and unspecified content about Covid-19. Moreover, social network websites, such as Facebook, Google, WhatsApp, and Microsoft, have pledged to work with governments to fight the spread of fake news. In the middle east, Saudi Arabia is one of the countries that is seriously fighting the spread of misleading information about the Covid-19 pandemic. The Public Prosecution in Saudi Arabia applied high penalties against fake news propagators with fines of up to 3 million SAR and imprisonment of up to 5 years.

Although the media paid considerable attention to studying the spread of fake news [8], most of these studies were conducted in western countries, with the English being the primary language of the studied samples. There are few empirical investigations of the diffusion of fake news in the Arab region. Cultural disparities are evident, highlighting the importance of researching the dissemination of fake news in many cultural contexts. In addition, studies of the diffusion of fake news during the time of pandemics and crises are currently limited. Therefore, this research aims to answer the following two questions: how to detect Arabic fake news about Covid-19 on Twitter through the use of rule-based systems, and what are the current trends of fake news diffusion during the pandemic in the Arab region?

This paper starts with reviewing the literature, followed by introducing the methodology used in this research. Data analysis results and discussion will then be presented. Finally, the conclusion, limitations, and future work will be presented.

II. LITERATURE REVIEW

A. Fake News

Since the inception of fake news, there has been no single definition of fake news [3]. According to [9], fake news is also defined as a news article intentionally written to deliver false information for a different purpose. Therefore, fake news is described in this study as manipulated false or misleading information to appear like real news for several purposes.

Several researchers were interested in studying the spread of fake news. For example, Allcott et al.[10] did a study to measure trends in the diffusion of misinformation circulating on social media from January 2015 to July 2018. The result from their research shows users' interactions with incorrect information rose steadily on both Facebook and Twitter up to the end of 2016. However, after one month, the interactions with false information dropped sharply on Facebook while rising on Twitter. This may result from a change in the Facebook platform after the 2016 elections to combat fake news. Another study was conducted by [2] to develop a method to overcome the spread of fake news in health during the current outbreak. The study focused on determining the type of false health information and used social impact in social media (SISM) methodology to analyze data. In addition, they selected Facebook, Twitter, and Reddit as social media channels for analysis. The study found that posts focused on fake health information are most aggressive. The spread of fake news during the current pandemic, i.e., Covid-19, made many individuals fearful and panicked. According to [2], psychological and neurological problems increased during the current pandemic, with fake news playing a significant role. In a recent study, researchers designed a dashboard to track misinformation on popular social media news sharing platforms (i.e., Twitter). To do so, they collected data from the platform beginning on March 1, 2020, and up to date [5]. This dashboard aims to fight false information and increase awareness about Covid-19.

Fake news can be detected by using three main methods. The first method relies on analyzing the news context where linguistic features are extracted and analyzed to identify fake news based on the writing styles that are commonly used in fake news [11, 12]. In addition, visual features can be used and analyzed to identify fake images [3]. The second method relies on analyzing social context where user-based, post-based, and networks-based segments are analyzed to determine fake news. User-based features such as users' profiles and characteristics can be analyzed to detect fake news through identifying the source of the fake news [12]. Post-based features mainly identify fake news by extracting and analyzing people's responses toward fake news [3]. Networks-based featured can remove by building specific networks among the users who published related posts [3]. The third and most recent one is using a knowledge-based context model, which aims to use external sources to check facts and identify fake news [13]. This last method is currently widely used because of its

accurate results and the increasing number of available high-quality fact-checking websites.

B. Text Mining and Text Classification

With the massive volume of data, companies and organizations started to use text mining techniques to improve their services, monitor brand reputation, gain a competitive advantage, and understand customers' behavior [14]. Text mining is used to extract hidden meaningful information from text [15, 16]. Moreover, text mining deals with unstructured data, while data mining deals with structured data [15, 17]. The process of mining text includes several stages: data gathering, data preparation, text transformation, feature selection, pattern selection, and evaluation [17]. Text mining involves a large set of algorithms and techniques for analyzing text, such as information retrieval, natural language processing, text summarization, classification (supervised learning), and clustering (unsupervised learning) [15]. These algorithms and techniques are being used in several contexts such as Business Intelligence, Customer care services, Knowledge management, Bioinformatics, Web Search Enhancement, and Risk Management [17].

In detecting fake news, the study [18] used semantic features and text mining to detect the spread of fake news in online articles. The used dataset was obtained from Kaggle about real-or-fake news. They applied five semantic features, including term frequency (TF), term frequency-inverse document frequency (TFIDF), bigrams, trigrams, quad-grams, and glove word embeddings along with Naive Bayes, random forest, and recurrent neural networks (RNN) classifiers. The study points out the bigram features with the random forest classifier achieved the best accuracy of 95.66%. So, text mining techniques are beneficial for finding misinformation. Another study was conducted by [19] to understand the impact of Covid-19 in Mexican society by using the text mining approach. The study extracted Twitter tweets about Covid-19 from the 13th to the 20th of March 2020, and the geo-localization of the retrieved tweets was Mexico City. The study found a positive correlation between the number of these tweets each day and the number of positive Covid-10 cases reported by the government on the same day. In addition, the people were fearful of health risks and economic crises that Covid-19 may cause. This required developing strategies to comfort the fear and panic associated with Covid-19.

Text classification is one of the essential methods in supervised learning [17]. It aims to assign classes or labels to texts, and it is used in many applications like image processing, document organization, etc. [15]. There are many learning algorithms used for text classification, such as Naïve Bayes Classifier, decision trees, Neural networks, and rule-based classifier. A team of researchers has developed a text-based algorithm, i.e., a supervised decision tree model, for analyzing customers' comments about a famous food brand on Twitter. The developed model predicted about 85% negative comments and 15% positive comments after analyzing 500 tweets [14]. A recent study applied text classification on Twitter data to analyze public fear sentiment during the Covid-19 pandemic [20]. Two machine learning (ML) classification methods were used, i.e., Naïve Bayes and logistic regression. Over nine hundred thousand tweets from February to March of 2020 were

analyzed using R software. The Naïve Bayes was able to classify public fear sentiment with a 91% accuracy rate, while the logistic regression was able to classify public fear sentiment with only a 74% accuracy rate. It is noted that the text classification helped in understanding the feeling of the public. Hence, it can be used to study social phenomena such as the spread of fake news.

1) *Rule-based classifier*: The rule-based classifier is one approach to text classification which uses a set of rules to separate the text into distinct groups [21]. Each of these rules contains an antecedent part, and a consequent part uses a series of "if-then" to represent them [22]. In [23] used a rule-based classifier with supervised learning to perform a Twitter sentiment analysis. Rules were built based on the occurrences of words related to emotion and opinion within the text. The study found that a rule-based classifier can improve the support vector machine SVM's predictions.

In summary, studies concerned with detecting fake news used either rule-based models or machine learning techniques to detect fake news. In this research, a rule-based classifier with a developed dictionary will be used to classify and detect fake news.

III. RESEARCH METHODOLOGY

Twitter data analysis has occupied a large volume of research in recent years, with many researchers relying on text mining techniques such as classification and clustering [16]. This study will develop an Arabic dictionary to detect fake news on social media (i.e., Twitter) and study the propagation of fake news in the Arab region. For our Arabic dictionary, we gathered fake news about the Covid-19 pandemic from various fact-checking websites, then built our dictionary around it. Our detection model is based on text classification that depends on rule-based systems to classify the tweets as fake or not fake news. The method of this study was divided into three steps: data collection, data preparation, and model building. Besides, we used python language (Jupyter notebook) to preprocess and analyze the dataset, Microsoft Excel to build the dictionary, and Power BI for data visualization.

A. Data Collection

We used a dataset of Arabic tweets on Covid-19 released on GitHub by Dr. Sultan Almujaivel¹ after getting his consent to use the data. The data were gathered between March 1, 2020, and April 30, 2020. To retrieve the dataset from Twitter, he used Arabic hashtags related to coronavirus disease to retrieve relevant tweets. Those hashtags are "#كورونا", "#covid-19", "#فيروس كورونا", "#coronaviruse", "#منع التجول", "#curfew", "#حظر التجول", "#curfew", "#كورونا_المستجد", "#novel_coronaviruse", "#العالم_في_مواجهة_كورونا", and "#world_facing_covid-19". The number of tweets in this dataset was 5015111 tweets.

B. Data Preparation

To clean and prepare the dataset for analysis, a series of processes were conducted. Noised tweets containing ads, coupons, and other irrelevant tweets were removed, as were old

tweets tweeted on a date other than the study's covered period. Arabic and English punctuation, numbers, hashtags, emoji, and empty tweets were removed too. Additionally, English letters, @username, and website links were replaced with an empty string. We also normalized the tweets in Arabic. Furthermore, missing values in the Location and Username columns were replaced with "undefine" values. As the data preparation finishes, the total number of tweets eligible for analysis was lowered to 4643425.

C. Model Building

To analyze the data, we performed two steps to building the model:

1) *The development of fake news dictionary*: The main goal of this study's dictionary development is to categorize Arabic tweets that contain fake news. Tweets that have spread in Arab communities are included in the dictionary. Initially, fake news was gathered from fact-checking websites as well as the fake news set published by [24] in GitLab². The fact-checking websites that were used in this study are Misbar³, Norumors⁴, AFPfact check⁵, Fatabyano⁶, Google fact check tools⁷. Duplicate news was checked and removed as it was gathered from various websites. A total of 212 pieces of fake news were discovered. Each piece of news was then tokenized, and the seven most important words were taken into account. One of these words has been identified as the primary key to fake news, with the remaining words serving as secondary keys. Table I shows two examples of fake news and how to extract the crucial words, where the "Fake" column represents the fake news, the "Key" column represents the fake news's primary key, and the reminder columns represent the fake news's secondary keys⁸. Following that, the fake news in the dictionary was classified into six categories, which corresponded to the categories identified by [24]. Table II. represents the categories and the number of relevant fake news categorized under each category in our dictionary. An example for each of the categories is presented in Table III.

2) *Rule-based model development*: The rule-based model was performed using two steps. First, the rule-based model will determine whether the primary key in the dictionary of fake news can be found in the text of the tweet. If the first condition is met, the model will check to see if at least two words from the dictionary's secondary words are found in the tweet's text. As a result, if the two conditions are met, the model will verify that the text does not contain any vocabularies that reject fake news, such as: ('false claim- ادعاء المعلومات- 'misinformation- ' لا صحه لذلك- ' زائف المغلوطة'). This is significant because some accounts on Twitter

² <https://gitlab.com/bigirqu/ArCOV-19>

³ MISBAR. 2020. Available: <https://misbar.com/>

⁴ RUMORS, N. 2020. Available: <http://norumors.net/>

⁵ ORGANIZATION, F. C. 2020. Available: <https://factcheck.afp.com/>

⁶ FATABYYANO. 2020. Available: <https://fatabyyano.net/>

⁷ TOOLS, G. F. C. 2020. Available:

<https://toolbox.google.com/factcheck/explorer>

⁸ The fake news dictionary we have developed is available on GitHub: <https://github.com/FatimahLAlotaibi/fake-news-dictionary>

¹ <https://github.com/salmujaivel>

fight and reject fake news. Therefore, by checking the existence of these vocabularies, one can verify that the tweet categorization as fake news is accurate. If the preceding conditions are met, the fake news category will be retrieved from a dictionary and assigned to the relevant tweet. Otherwise, the tweet will be categorized as not-fake news. Fig. 1 depicts the code used to determine whether a tweet contains fake news and to assign tweets to the appropriate category. The second step aims to label tweets based on the category, with label =1 indicating that the tweet contains fake news and label =0 indicating that the tweet does not contain fake news. To do so, we look to see if the category contains a "not" value, which indicates that the label is equal to zero; otherwise, the label is equal to one. Fig. 2 depicts the code that was used to assign the tweets to the label.

TABLE I. EXAMPLES OF FAKE NEWS IN THE DICTIONARY

Fake	Key	S1	S2	S3	S4	S5	S6
ثلاثة ادوية تسبب الموت السريع عند الإصابة بفيروس كورونا أو الاشتباه به	ادويه	ثلاث	أصابه	السريع	سبب	تؤدي	اشتباه
Three medicines cause rapid death when infected with coronavirus or suspected of it.	Medicines	Three	Infection	Rapid	Cause	lead to	Suspected
رمي جثث موتى كورونا في البحر النجر بالمكسيك	جثث	بحر	حذف	رمي	مكسيك	قتلى	موتى
Throwing corpses of people dead by Corona in the sea in Mexico.	Corpses	Sea	Disposal	Throw	Mexico	Dead	Dead

TABLE II. CATEGORIES OF FAKE NEWS IN THE DICTIONARY

Category	Entertainment	Health	Politics	Religious	Social	Sports	Total
Number of fake news	13	41	29	33	89	7	212

TABLE III. EXAMPLE OF EACH CATEGORY OF FAKE NEWS

Fake (Arabic)	Fake (English)	Category
الممثلة الإسبانية ألبا فلوريس التي اشتهرت بدور نيروبي توفيت بسبب فيروس كورونا المستجد	Spanish actress "Alba Floris" who is known as "Nairobi" was death by Covid-19.	Entertainment
ثلاثة ادوية تسبب الموت السريع عند الإصابة بفيروس كورونا أو الاشتباه به	Three medicines cause rapid death when infected with coronavirus or suspected of it.	Health
سياسي فرنسي يرمي علم الاتحاد الأوروبي بسبب تخليه عن الدول الأعضاء خلال أزمة كورونا المستجد	French politician throws the European Union flag due to its abandonment of the member states during the Corona pandemic.	Politics
أول صلاة جمعة في الصين بعد إغلاقها أعوام واعوام	The first Aljumuah prayer in China after being close for years	Religious
وزارة الداخلية تحذر فيه من سرقات عديدة حصلت بسبب عصابات تدعي أن هدفهم تعقيم البيوت من فيروس كورونا	The interior ministry warns of numerous thefts that happened due to gangs claiming their goal is home sterilization from coronavirus.	Social
إصابة حارس مرمى نادي ريال مدريد ومنتخب بلجيكا، تيبو كورتوا، بفايروس كورونا.	The goalkeeper of Real Madrid and the Belgian national team "Tibo Coroto" has been infected with the coronavirus.	Sports

```

category='not'
count=0
i=0
while i<len(s):
    if s[key][i] in (text):# if basic word from dictionary in tweet
        if s[s1][i] in (text):# if secondary word from dictionary in tweet
            count=count+1
        if s[s2][i] in (text):
            count=count+1
        if s[s3][i] in (text):
            count=count+1
        if s[s4][i] in (text):
            count=count+1
        if s[s5][i] in (text):
            count=count+1
        if s[s6][i] in (text):
            count=count+1
    if count>1:
        category=s[Category][i] # assign category to tweet
        for x in R: # R is list of word that reject fake news
            if x in (text):
                category="not"
                break
        break
    count=0
    i=i+1
    
```

Fig. 1. Code to Check if the Tweets are Fake News or Not.

```

data_df1["model-label"]=0
i=0
while i<len(data_df1):
    if data_df1["Category"][i]!="not":
        data_df1["model-label"][i]=1
    i=i+1
    
```

Fig. 2. Code Assigning the Tweets to the Matching Label.

IV. RULE-BASED MODEL EVALUATION

To evaluate and confirm the proposed model, a balanced sample of 2000 tweets was extracted at random. The tweets in the sample were labeled manually, with label=1 indicating fake news and label=0 indicating not-fake news (i.e., real news or others). In this sample, there were 997 tweets labeled as fake news and 1003 tweets labeled as not fake news. The sample was then divided into 70 percent (1400 tweets) for training our model and the remaining 30 percent (600 tweets) for testing. The performance of the proposed model was assessed using accuracy, precision, and recall. These three metrics are measured by calculating the number of correct positive predictions (TP), the number of correct negative predictions (TN), the number of incorrect positive predictions (FP), and the number of incorrect negative predictions (FN) [9], as shown respectively in equations 1, 2, and 3.

$$\text{Accuracy} = \left(\frac{(|TP|+|TN|)}{(|TP|+|TN|+|FP|+|FN|)} \right) \quad (1)$$

$$\text{Precision} = \left(\frac{(|TP|)}{(|TP|+|FP|)} \right) \quad (2)$$

$$\text{Recall} = \left(\frac{(|TP|)}{(|TP|+|FN|)} \right) \quad (3)$$

Table IV displays the precision, recall, and accuracy of our model when applied to training and testing data in this sample. Our model's prediction level is adequate based on these results. As a result, we may use the model to examine our data.

TABLE IV. ACCURACY, PRECISION, AND RECALL FOR VALIDATING THE RULE-BASED MODEL

	Training Data	Testing Data
Accuracy	79.7%	78.1%
Precision	71.2%	70%
Recall	99%	98%

V. MODEL EVALUATION USING FUTURE DATASET

To validate the proposed model, the researchers obtained 545 tweets at random as future data. The tweets were then manually labeled. When the model was applied to these data, the accuracy was 95.9 percent, indicating that the model was acceptable and applicable.

VI. RESULTS AND DISCUSSION

A. Fake news Propagation Trends across the Arab Region

After ensuring that our model had an acceptable level of accuracy, it was applied to the dataset. 26006 tweets were

labeled as fake news by the model. According to an analysis of the source of these tweets, only 13491 fake news tweets have location details. This is because location data on Twitter is provided only if the user has a geo-enabled feature or mentions a valid location in his public profile [5]. The analysis of the location of fake news tweets reveals that the majority of fake news tweets originated in Saudi Arabia, Egypt, Kuwait, Qatar, Yemen, United Arab Emirates, Iraq, Palestine, Lebanon, Jordan, Oman, Algeria, Morocco, Libya, Sudan, Syria, Bahrain, and Tunisia respectively (see Fig. 3). Given that the majority of the analyzed tweets originated from Saudi Arabia and Egypt, it was expected that a high number of fake news tweets are being exchanged in these countries. This is also consistent with recent statistics showing that Saudi Arabia and Egypt are ranked 8th and 18th in the world in terms of the number of Twitter users, respectively⁹. However, the high number of fake news tweets from Kuwait, which is so close to Egypt, is surprising. Kuwait is a small country with a population of only 4.4 million¹⁰ people. Nevertheless, about 99.5%¹¹ of its population are current Internet users, and more than half of them are Twitter users. Therefore, this may have affected the spread of fake news in Kuwait.

As previously stated, the model was trained to categorize fake news tweets into six categories based on the general theme of the tweet. Table V shows the number of tweets in each category of fake news.

B. Fake News Word Cloud

After extracting the fake news tweets from the dataset, we used a word cloud to analyze the frequency of the words in these tweets. However, it was decided to exclude two words (corona "كورونا"; and virus "فيروس") because they were found to harm the word cloud's results. Fig. 4 depicts the word cloud results for the most common words found in fake news tweets. The most frequent tweeted fake news was related to the daily use of hot steam inhalation to 'kill' the Covid-19 coronavirus.

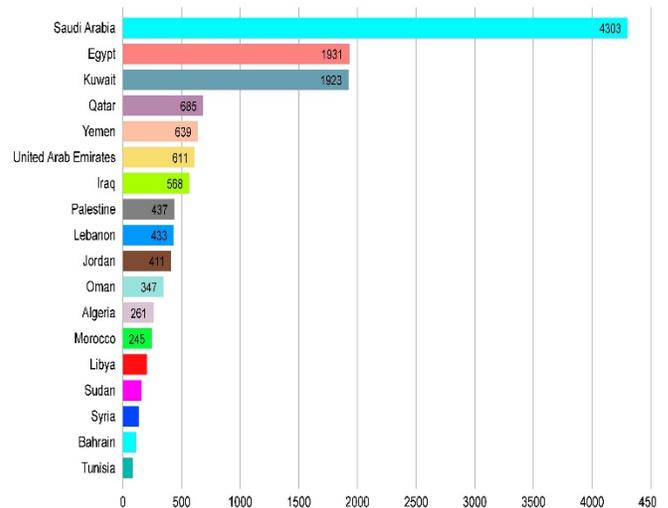


Fig. 3. Number of Fake News in Arab Countries.

⁹ <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/>

¹⁰ Central Statistical Bureau, <https://www.csb.gov.kw>

¹¹ The World Bank Open Data, <https://data.worldbank.org/>

TABLE V. CATEGORIES OF FAKE NEWS TWEETS AND NUMBER OF RELEVANT TWEETS

Category	Number
Entertainment	629
Health	3911
Politics	4032
Religious	6338
Social	9781
Sports	1315
Total	26006

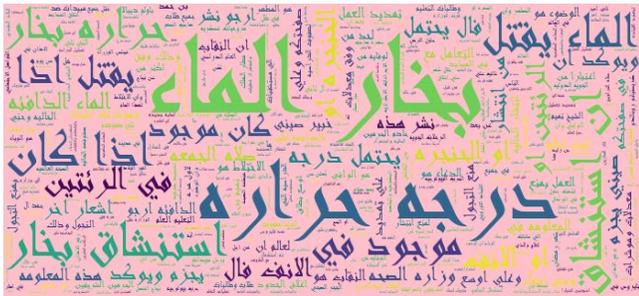


Fig. 4. Word Cloud for Fake News.

C. Analysis of the Fake News across the Dates

The dataset for our study was obtained from Twitter between March and April. After analyzing the data, we discovered that the spread of fake news tweets was much higher in March than in April. Fig. 5 shows that approximately 63.06% of fake news tweets were posted in March, while only 36.94% were posted in April indicating a decrease in the number of fake news tweets in April. These findings show that ambiguity surrounding the coronavirus and people's feeling of fear may have increased fake news propagation at the beginning of this crisis. Therefore, as users' awareness and knowledge of Covid-19 increases, the spread of fake news tweets decreases.

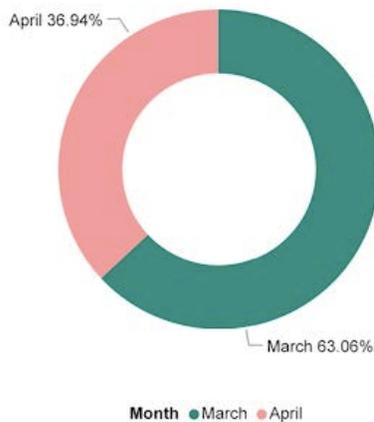


Fig. 5. Distribution of Fake News Tweets across Categories.

Because the number of fake news tweets in March is higher than in April, we attempted to improve our understanding of the spread of fake news tweets in March by analyzing the distribution of fake news during the month's days. Fig. 6 shows that the 14th, 15th, 16th, and 25th of March saw the most incredible spread of fake news tweets in the Arab region. On or around March 15th, most Arab countries suspended in-class study and switched to online study for both schools and universities. Egypt, for example, temporarily suspended traditional education on March 15th, which explains why the 14th, 15th, and 16th of March have the highest number of fake news tweets. Furthermore, the partial curfew in Saudi Arabia began on March 23rd and in Egypt on March 25th, which could explain the rise in fake news on March 25th. This demonstrates that as more information and knowledge about Covid-19 become available over time, user awareness of the current pandemic grows, while the number of spread fake news tweets decreases.

As previously observed, Saudi Arabia, Egypt, and Kuwait had the highest number of fake news tweets. As a result, we attempted to comprehend the distribution of fake news during March and April. Fig. 7 shows that the number of fake news tweets in Egypt and Kuwait decreased by roughly half between March and April, but not significantly in Saudi Arabia. It is plausible that the total curfew imposed by Saudi Arabia in major cities on April 6th contributed to the continuation of fake news.

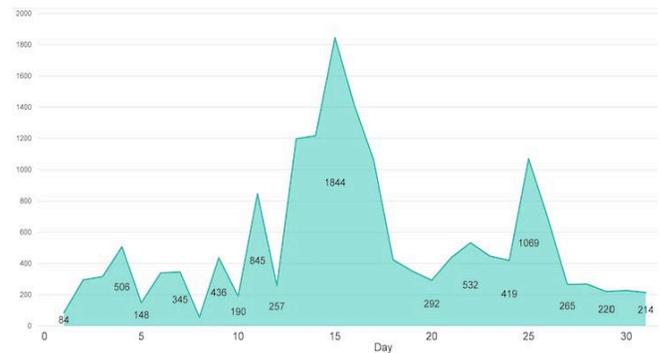


Fig. 6. Spread Fake News Tweets during March.

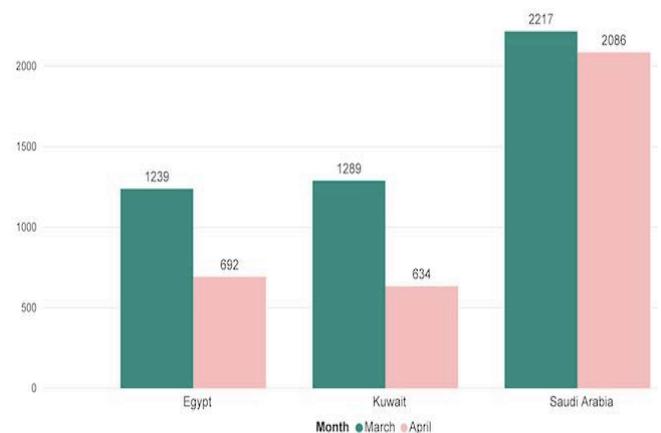


Fig. 7. Spread of Fake News Tweets by Months in the Top Three Arabic Countries

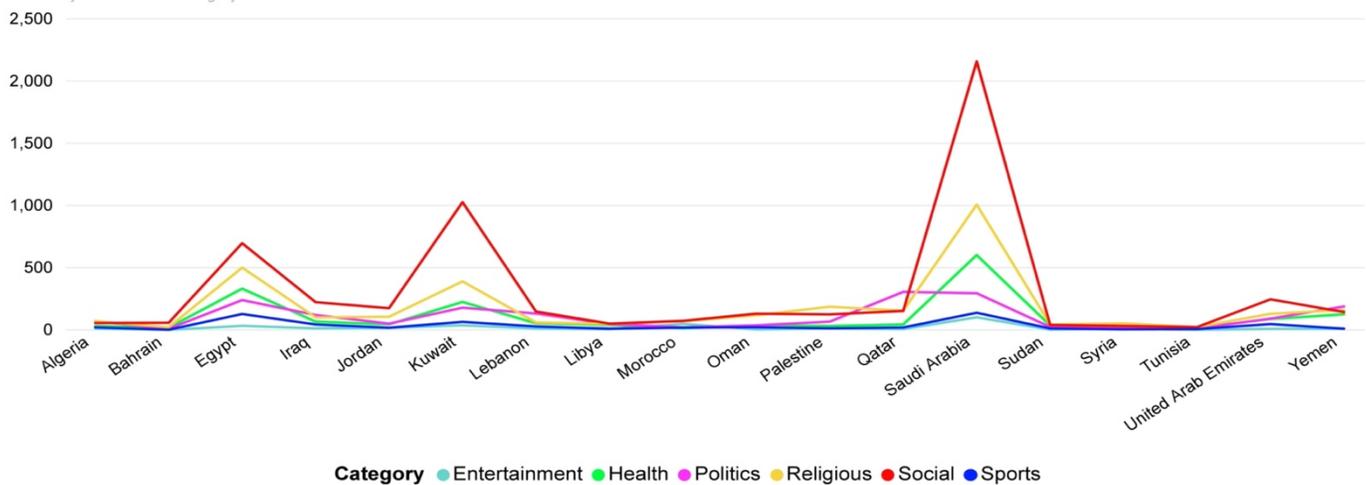


Fig. 8. Categories of Fake News across Arab Countries.

D. Distribution of Fake News Categories across Arab Countries

Fig. 8 depicts the various categories of fake news tweets in Arab countries. Fake news tweets about social issues are the most prevalent in all Arab countries except Palestine, Qatar, Yemen, and Algeria. In Palestine, fake news tweets were mostly about religion, whereas in Qatar and Yemen, fake news tweets were mainly about politics. In Algeria, fake news tweets were mostly about religious or political issues. Fake news related to entertainment, on the other hand, was the lowest across all Arab countries, except Morocco, which has a deficient number of fake news tweets related to health topics.

VII. CONCLUSION

This study was carried out to understand better the spread of the fake news phenomenon in the Arab region during Covid-19. We created an Arabic dictionary and used text classification, i.e., a rule-based system, to analyze the spread of fake news in the Arab region using a secondary dataset to detect fake news. We discovered that the number of fake news spreads on Twitter was much higher in March than in April and that dates with significant response measures taken to control the spread of Covid-19 had the highest number of fake news spreads. This figure usually falls a few days after the measurement is taken. This demonstrates that at the start of new changes, people panic, mainly due to a lack of information about what will happen next, and that as more information becomes available, users adapt to the new norms and stop spreading such fake news.

VIII. LIMITATION AND FUTURE RESEARCH

This study has some limitations. The primary limitation is the difficulty of processing Arabic. The Arabic language, unlike other languages, has distinct writing principles such as writing from the right to the left side, the absence of capital letters, and distinct grammatical rules for detecting entities, acronyms, and abbreviations. The presence of many dialects in the Arab region, the use of slang words and colloquial terms, and many spelling errors are also significant challenges when analyzing the Arabic language. The tweets dataset used in this

study came from various Arabic countries, and the tweets were written in a variety of Arabic dialects, making it difficult to process and analyze texts using stemming. To improve the efficiency of the analysis, future work should focus on developing the stemming process for the Arabic language. Additionally, because new fake news is created periodically, the developed fake news dictionary must be updated regularly. In addition, the result of the proposed model in this study will be compared in the future with the results of the machine learning models to extra validate and assess the performance of the proposed model. Also, an additional dataset from other social media platforms, such as Facebook, will be used in the future to see if the same trends of fake news dissemination are observed across platforms.

ACKNOWLEDGMENT

The authors would like to thank Dr. Sultan Almujaivel, who works at King Saud University, for providing them with the dataset of Covid-19 Arabic tweets used in this study.

REFERENCE

- [1] A. Al-Rawi, Gatekeeping fake news discourses on mainstream media versus social media. *Social Science Computer Review*, 2019. 37(6): p. 687-704.
- [2] C.M. Pulido, L. Ruiz-Eugenio, G. Redondo-Sama, and B. Villarejo-Carballido, A New Application of Social Impact in Social Media for Overcoming Fake News in Health. *International journal of environmental research and public health*, 2020. 17(7): p. 2430.
- [3] K. Shu, et al., Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 2017. 19(1): p. 22-36.
- [4] A. Spinelli and G. Pellino, COVID - 19 pandemic: perspectives on an unfolding crisis. *The British journal of surgery*, 2020.
- [5] K. Sharma, et al., Covid-19 on social media: Analyzing misinformation in twitter conversations. *arXiv preprint arXiv:2003.12309*, 2020.
- [6] M.K. Elhadad, K.F. Li, and F. Gebali. COVID-19-FAKES: A twitter (Arabic/English) dataset for detecting misleading information on COVID-19. in *International Conference on Intelligent Networking and Collaborative Systems*. 2020. Springer.
- [7] Y. Chen, et al., A discussion of irrational stockpiling behaviour during crisis. *Journal of Safety Science and Resilience*, 2020. 1(1): p. 57-58.
- [8] C. Silverman, This analysis shows how viral fake election news stories outperformed real news on facebook, 2016. URL <https://www.buzzfeed>.

- com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook, 2016.
- [9] S. Yang, et al. Unsupervised fake news detection on social media: A generative approach. in Proceedings of the AAAI Conference on Artificial Intelligence. 2019.
- [10] H. Allcott, M. Gentzkow, and C. Yu, Trends in the diffusion of misinformation on social media. *Research & Politics*, 2019. 6(2): p. 2053168019848554.
- [11] Y. Chen, N.J. Conroy, and V.L. Rubin, Misleading Online Content: Recognizing Clickbait as "False News", in Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection. 2015, Association for Computing Machinery: Seattle, Washington, USA. p. 15–19.
- [12] C. Castillo, M. Mendoza, and B. Poblete, Information credibility on twitter, in Proceedings of the 20th international conference on World wide web. 2011, Association for Computing Machinery: Hyderabad, India. p. 675–684.
- [13] A. Vlachos and S. Riedel. Fact checking: Task definition and dataset construction. in Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. 2014.
- [14] A.S. Halibas, A.S. Shaffi, and M.A.K.V. Mohamed. Application of text classification and clustering of Twitter data for business analytics. in 2018 Majan International Conference (MIC). 2018. IEEE.
- [15] M. Allahyari, et al., A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*, 2017.
- [16] S.A. Salloum, M. Al-Emran, A.A. Monem, and K. Shaalan, A survey of text mining in social media: facebook and twitter perspectives. *Adv. Sci. Technol. Eng. Syst. J.*, 2017. 2(1): p. 127-133.
- [17] S. Sheela and T. Bharathi, Analyzing Different Approaches of Text Mining Techniques and Applications. *International Journal of Computer Science Trends and Technology*, 2018. 6(4): p. 23-29.
- [18] P. Bharadwaj and Z. Shao, Fake news detection with semantic features and text mining. *International Journal on Natural Language Computing (IJNLC) Vol*, 2019. 8.
- [19] J.E.C. Saire and A. Pineda-Briseno, Text mining approach to analyze coronavirus impact: Mexico city as case of study. *medRxiv*, 2020.
- [20] J. Samuel, et al., Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, 2020. 11(6): p. 314.
- [21] M. Thangaraj and C. Vijayalakshmi, Performance study on rule-based classification techniques across multiple database relations. *International Journal of Applied Information Systems*, 2013. 5(4): p. 1-7.
- [22] U.A. Siddiqua, T. Ahsan, and A.N. Chy. Combining a rule-based classifier with ensemble of feature sets and machine learning techniques for sentiment analysis on microblog. in 2016 19th International Conference on Computer and Information Technology (ICIT). 2016. IEEE.
- [23] P. Chikersal, S. Poria, and E. Cambria. SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). 2015.
- [24] F. Haouari, M. Hasanain, R. Suwaileh, and T. Elsayed, ArCov-19: The First Arabic COVID-19 Twitter Database with Propagation Networks. 2020.

Hybrid Deep Neural Network Model for Detection of Security Attacks in IoT Enabled Environment

Amit Sagu, Nasib Singh Gill, Preeti Gulia
Department of Computer Science and Applications
Maharshi Dayanand University
Rohtak, Haryana, India

Abstract—The extensive use of Internet of Things (IoT) appliances has greatly contributed in the growth of smart cities. Moreover, the smart city deploys IoT-enabled applications, communications, and technologies to improve the quality of life, people's wellbeing, quality of services for the service providers and increase the operational efficiency. Nevertheless, the expansion of smart city network has become the utmost hazard due to increased cyber security attacks and threats. Consequently, it is more significant to develop the system models for preventing the attacks and also to protect the IoT devices from hazards. This paper aims to present a novel deep hybrid attack detection method. The input data is subjected for preprocessing phase. Here, data normalization process is carried out. From the preprocessed data, the statistical and higher order statistical features are extracted. Finally, the extracted features are subjected to hybrid deep learning model for detecting the presence of attack. The proposed hybrid classifier combines the models like Convolution Neural Network (CNN) and Deep Belief Network (DBN). To make the detection more precise and accurate, the training of CNN and DBN is carried out by using Seagull Adopted Elephant Herding optimization (SAEHO) model by tuning the optimal weights.

Keywords—Internet of things; deep learning; optimization; convolutional neural network; security attack detection

I. INTRODUCTION

IoT is an integration of services, people, interconnected entities, and physical infrastructure that process the information [1]. Moreover, the IoT systems are dynamically distribute, edge based computational resources and allocations of information. IoT devices communicate with one another by wireless communication systems and transfer the information to a centralized model [2][3]. IoT is one of the interrelated models that supports seamless information among the devices (e.g.), automotive sensors, environmental sensor, industrial robots, road-side sensors, surveillance devices, medical devices and smart home sensors. The sum figure of the linked IoT devices has touched the usage of 27 billion in 2017. IoT devices used various technologies, service types, and protocols [4]. Consequently, it seems more complex to maintain the upcoming IoT framework as it leads to unwanted vulnerability in the environment. The cyber-attack could access the details in an illegal manner regarding each activity of citizens without the user's knowledge or can reconfigure the devices with the unsecured settings [5][6].

The risk rendering through these attacks may affect the protection of IoT networks and the entire eco-system such as

applications, web-sites, servers and social networks, through malicious smart device known as botnet (i.e.), robot networks. Also, a communication channels or single component in IoT-based systems can be compromised by paralyzing the complete or part of Internet network [7][8][9]. Hence, the standard attack detection model is required, which could analyze the behavior of attacks in network. The rising of deep learning (DL) has alleviated the limitations of the conventional machine learning (ML) schemes due to the combined implementation of classifier and feature extraction, and its strong representative ability [10] [11][12] Further, DL model is used for avoiding the overhead of manual selection of features, and that is an essential section for traditional classification systems [13][14].

Several researchers have used the DL tools for solving the problems related to the communication which is progressively being carried out [15] [16]. Particularly, DBN is implemented based on the AMC scheme through SCF feature; however, it attains the restricted classification outcomes due to inadequate ability. Furthermore, an unsorted DNN is used for identifying the signal modulation systems with less computational complication [17] [18] [19] [20][21]. Still, the deficiencies of the convolutional operation make it more complex for extracting the high-dimensional features. The IoT devices creates large amount of data. Moreover, the ML [22][23] pipelines has performed the process of data collection, feature extraction, and binary classification in many systems or models for the detection of IoT traffic. Several ML algorithms [24][25] [26][27] including NNs[28][29][30], BNs, EL, clustering, FS, SVMs, and DTs are used for IoT attack detection with great impact. Recently, DL model is used for detecting the anomalous behavior in the IoT field [31][32][33].

The key contributions of the proposed model are given below:

- Introduces the Hybrid model for detection of attack in IoT.
- Proposed the Seagull Adopted Elephant Herding Optimization (SAEHO) algorithm for training the hybrid system through tuning the optimal weights.

In this paper, the literature review on attack detection in IoT is given in Section II. Overall description of the adopted attack detection model is determined in Section III. Pre-processing and feature extraction phase are described in

Section IV. Section V describes attack detection by proposed hybrid deep learning model. Section VI depicts the proposed seagull adopted elephant herding optimization algorithm for optimal training of hybrid model via tuning the weights. Section VII specifies the result and discussion. At the end, the conclusion of this paper is depicted in Section VIII.

II. LITERATURE REVIEW

A. Related Works

In [34] Li, et al., (2019) has introduced the information security approach of block chain on the basis of intrusion detection technique in the IoT. Moreover, the intrusion detection technology was used for analyzing the recognition technology on the basics of dissimilar systems, and the security of block chain information. From hacker attack, the intrusion detection model was one of the security technologies for protecting the network resources. IDS were more beneficial enhancement to the firewall that would assist the network approach for enhancing the integrity of the information security framework and detecting the attacks quickly. Finally, the proposed intrusion detection technique was used for the block chain information security system, and the experimental outcomes have shown better fault tolerance and higher detection efficiency.

In [35] Boubeta, et al., (2020) has proposed an intelligent architecture which combined ML paradigm and the CEP technology for detecting various categories of security attacks in real time IoT. Additionally, the proposed architecture was accomplished for managing the event patterns easily and the conditions depend on values attained via ML models. Moreover, an automatic code generation and a model-driven graphical device were provided for pattern definition in security attack and it hides all the complication attained from execution information of domain experts. The simulation outcome of the adopted model has demonstrated better performance than other schemes.

In [36] Marcos, et al., (2020) have adopted a near real-time SDN security model in which it secures the basis of SDN controller besides the traffic destruction and avoids the DDoS attacks in the source-end network. Further, the CNN for DDoS detection was tested and applied, and determined the system alleviate the identified attacks. A GT based technique was used to mitigate the attack in which it optimized the packet discard rate and concern within the SDN's central controller. At the end, the experimental results of the presented SDN security scheme have shown better outcomes against next-generation DDoS attacks.

In [37] Mabodi, et al., (2020) have determined a hybrid system based on the cryptographic authentication. In addition, the adopted model includes four stages like gray hole attack discovery, testing the routes, the malicious attack removal procedure in MTISS-IoT, and the IoT identifying node trust. The adopted system was assessed via extensive simulations that were done in the NS-3 tool. At the end, the experimental results of four circumstances have determined that the MTISS-IoT model has shown better FPR, FNR, and detection rate than other models.

In [38] Chunsheng, et al., (2020) have implemented a MMFN for identifying the signal modulations via a new feature known as PCCs. Furthermore, a PCCP was implemented for converting the raw modulated signals into PCCs, and it was the inputs given to MMFN. The multi-module fusion model was proposed in MMFN for acquiring the higher representation capability. Moreover, the characterization module was implemented for balancing the tradeoff among the dimensions of the extracted features and the number of parameters. At the end, the experimental results of the presented MMFN approach have achieved 90% accuracy at 1 dB SNR and superior classification performance.

In [39] Mohammed, et al., (2020) has discussed the IoT-ED possibility with implanted HT which provides serious privacy, security, and available issues to the IoT based HAN. Moreover, the traditional network attack detection models have worked the network protocol layers, while the IoT-ED with HT leads to the demonstration of attack at the firmware or/and physical level. The adopted model was used for identifying the multiple attacks and differentiated the various attack types. Further, the IoT-ED behaviors have been studied for 5 various random attacks that includes the DoS, impersonation attacks, power depletion, ARQ, and covert channel. The adopted method could distinguish with 92% accuracy for all the attacks simultaneously.

In [40] Sahay, et al., (2020) have determined a layered scheme of IoT routing security for analyzing the susceptibility linked with every phase of the routing method. The adopted system has explored the leverage of inherent features in blockchain for enhancing the security in IoT-LLNs. Moreover, the blockchain network operated as a protected data link among the attack detection mechanism and the IoT-LLN to enhance the outcomes of XGBoost algorithm. Finally, the blockchain-based model was implemented with elegant contract to generate the real-time alerts for identifying the sensor nodes.

In [41] Alabady, et al., (2020) have introduced a novel security system in the IoT era for cooperative virtual networks. The proposed model has determined the attacks and risks in switches, network security vulnerabilities, threats, routers and firewalls, along with a policy for mitigating those risks. The adopted method has offered the basics of secure networking scheme that includes router, firewall, VLAN technology and AAA server. At last, the simulation results of the adopted approach have demonstrated an effective security execution with excellent network services and speed.

B. Review

Table I shows the review on attack detection system in IoT. Originally, the Mapping UML model was determined in [41] that presents higher detection efficiency, fault tolerance and better accuracy; however, the data selection sensor technique was not incorporated into the computer environment. Moreover, the CEP and ML models were deployed in [24] that provide better precision, higher recall, and maximum F1 score. Nevertheless, more event patterns were not defined in the proposed model for detecting other types of attacks. CNN model was exploited in [42] that offer

higher accuracy, improved precision rate, and maximum recall, but need to maximize the host count in the simulated SDN environment. Likewise, MTISS-IoT model was exploited in [23], which offers better FPR, low FNR and maximum detection rate. However, the firefly optimization was not used in the proposed work to lower consumption energy and malicious attacks on the IoT. MMFN method was exploited in [29] that have robustness, higher classification accuracy, and strong characterization ability; however, the small-scale data-driven DL- AMC model with less training time was needed for training the neural network (NN). In addition, an IoT-based HAN model was determined in [28], which offers better accuracy, reduced false positives, and high precision. However, the proposed work needs to suggest the moving data process nearer to the network edge. XGBoost Classifier was suggested in [33] that offers secured network, maximum accuracy, higher recall and improved operational efficiency. However, need to investigate an efficient mechanism to address and analyze the challenges. Finally, the VLAN was introduced in [35], that offers effective security execution, best network speed and services, but the VLAN technology were not utilized in the LAN environment. Thus, the challenges have to be taken into account based on attack detection method in IoT in the present work efficiently.

III. SYSTEM MODEL OF INTRUSION DETECTION ON IOT FRAMEWORK

The IoT plays significant role in the information age, and it is a significant component of the novel information technology. Moreover, the IoT server is the functional core of the entire IoT business scheme. The essential functions of terminal sensor processing, data collection and return the processing outcomes are all designed through the server. Further, the security vital in cyber life as it relies with great advancement of IoT techniques. In addition, the IDS are the protector for the Internet servers. Fig. 1 indicates the circumstances in which the IDS are concerned in the IoT network. Many of the IoT devices and IoT servers are exposed directly to the public Internet due to the feature of remote control. In addition, the attackers would capture the vulnerabilities for intruding the IoT servers. However, the IDS are extremely needed for protecting and detecting the IoT servers from the attackers. The IDS usage would protect the terminal users and also protect the service providers from the hazards on the Internet. The security protections are not fully achieved in the IoT application as it reduces the attack plane. The lowering of the attack plane is limited extremely, and intruders may find the path to crack the assured node in the network. This work seeks the strategy of deep learning concept in the intrusion detection system. Fig. 1 illustrates the IoT framework.

TABLE I. REVIEW ON TRADITIONAL ATTACK DETECTION MODEL IN IOT: FEATURES AND CHALLENGE

Author [citation]	Adopted scheme	Features	Challenges
Li <i>et al.</i> [34]	Mapping UML model	✓ Higher detection efficiency ✓ Fault tolerance ✓ Better accuracy	➤ The data selection sensor technique was not incorporated into the computer environment.
Boubeta, <i>et al.</i> [35]	CEP and ML models	✓ Better precision ✓ Higher recall ✓ Maximum F1 score	➤ More event patterns were not defined in the proposed model for detecting other types of attacks.
Marcos <i>et al.</i> [36]	CNN model	✓ Higher accuracy ✓ Improved precision rate ✓ Maximum recall	➤ Need to maximize the host count in the simulated SDN environment.
Mabodi <i>et al.</i> [37]	MTISS-IoT model	✓ Better FPR ✓ Low FNR ✓ Maximum detection rate	➤ The firefly optimization was not used in the proposed work to lower consumption energy and malicious attacks on the IoT.
Sai <i>et al.</i> [38]	MMFN method	✓ Robustness, ✓ Higher classification accuracy ✓ Strong characterization ability.	➤ The small-scale data-driven DL- AMC model with less training time was needed for training the NN.
Mohammed <i>et al.</i> [39]	IoT-based HAN model	✓ Better accuracy ✓ Reduced false positives ✓ High precision	➤ The proposed work needs to suggest the moving data process close to the network edge
Sahay <i>et al.</i> [40]	XGBoost Classifier	✓ Secured network ✓ Maximum accuracy ✓ Higher recall ✓ Improved operational efficiency	➤ Need to investigate an efficient mechanism to address and analyze the challenges.
Alabady <i>et al.</i> [41]	VLAN	✓ Effective security execution ✓ Best network speed and services	➤ The VLAN technology were not utilized in the LAN environment.

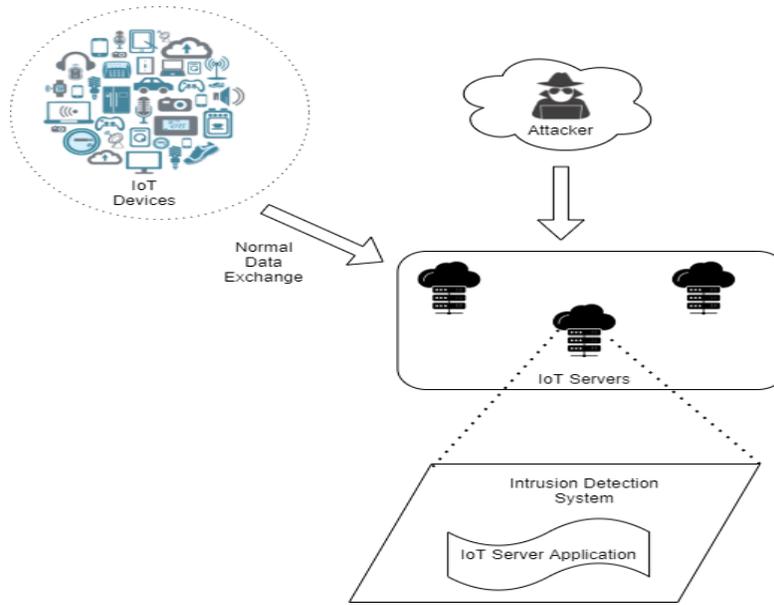


Fig. 1. IoT Framework.

IV. OVERALL DESCRIPTION OF THE ADOPTED ATTACK DETECTION MODEL IN IoT

This proposal intends to introduce a novel deep hybrid attack detection system that consists of three phases: “(i) preprocessing (ii) Feature extraction (iii) Classification”. Originally, the input data is preprocessed under data normalization process. Subsequently, the preprocessed data is subjected to the feature extraction stage, where the higher order statistical features and statistical features are extracted. Moreover, the statistical features include mean, median, SD, mode, HM, RMS, peak amplitude and pitch angle; and the higher order statistical features include kurtosis, skewness,

energy, entropy, mean frequency, and percentile are extracted. Moreover, the extracted features are provided as the input to the detection phase with hybrid model that combines the models like CNN and DBN. It is obvious that the detection model must be trained in a proper manner, such that the detection accuracy increases in this way. To make this possible, this work adhere the utilization of optimization logic that could make the training process more optimal. Thereby, the weights of both the CNN and DBN are optimally tuned by a new SAEHO algorithm. This is the proposed hybrid algorithm, which combine the logic of both the EHO and SOA algorithm. Fig. 2 illustrates the architecture of proposed detection system.

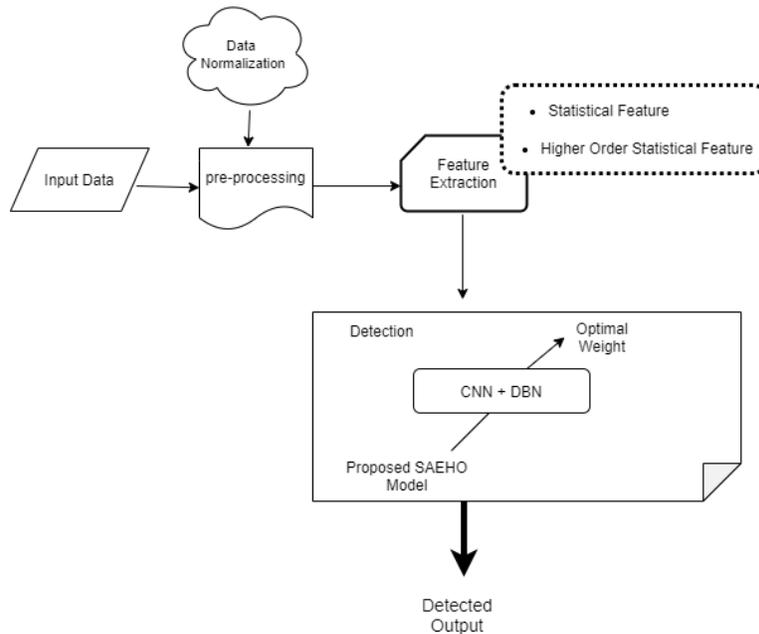


Fig. 2. Overall Framework of the Proposed Model.

V. ATTACK DETECTION BY PROPOSED HYBRID DEEP LEARNING MODEL

The proposed hybrid model follows the parallel execution of both the models with the extracted feature set, and finally averages the outcomes obtained from each model, which is considered as the final detection results.

A. Optimized CNN Model

The extracted features FE are provided as input to optimized CNN [43]. Convolutional Networks are the trainable multistage framework that includes numerous stages. The input and output of every stage are group of arrays recognized as feature maps. In addition, the well-recognized classifier is CNN that consists of 3 layers like “fully connected layer, pooling layer, and convolution layers”. Furthermore, the convolution layer contains of several convolution kernels. The entire feature map was determined by the numerous kernels. Moreover, the s^{th} layers matched to x^{th} feature map and feature values in the place (e, g) is denoted as $H_{e,g,x}^s$, and it is given in Eq. (1). Similarly, the x^{th} filter value is provided in the s^{th} layer. Consequently, the optimal tuning of the weight is performed using the adopted SAEHO scheme. The linked input patches in the x^{th} layer at location (e, g) are determined using $L_{e,g}^s$. The non-linearity is attained through the activation function which predicts the nonlinear features of multi-layer networks. Moreover, $(A_{e,g,x}^s)$ and $A(\bullet)$ are defined in Eq. (2). Even though, the shift-variance in the pooling layer is deployed through minimizing the resolution of feature maps as given in Eq. (3). $pool(\)$ and Local neighbourhood for every feature map $(A_{e,g,x}^s)$ at (e, g) is portrayed as $I_{e,g}$.

$$H_{e,g,x}^s = W_l^{sT} L_{e,g}^s + B_l^s \quad (1)$$

$$A_{e,g,x}^s = A(H_{e,g,x}^s) \quad (2)$$

$$O_{e,g,x}^s = pool(A_{e,g,x}^s), \forall (\hat{e}, \hat{r}) \in I_{e,g} \quad (3)$$

Eq. (4) determines the loss function in CNN. The constraints (ζ) of CNN are associated to the required IO input-output relation, and it is given as $\{(V^{(t)}, U^{(t)}); t \in [1, \dots, IO]\}$.

$$Loss = \frac{1}{Num} \sum_{t=1}^{ms} \hat{P}(\zeta; U^{(t)}, OUT^{(t)}) \quad (4)$$

Pooling layer: “In CNN, the pooling layer has performed the processes of down sampling with the resultant attained from the convolutional layers. Further, the 2 renowned pooling types such as max pooling and average pooling are used. The max pooling has attained the higher value; but the average value is observed in the average pooling”.

Fully connected layer: It works within the flattened inputs. In general, the results attained from the pooling layer are given as the input of fully connected layer and thus the inputs are connected to all layers. In the CNN structure, the fully connected layer occurs at its edges. The output of CNN is denoted as CL_{CNN} .

B. DBN based Attack Detection

In 1986, Smolensky implemented DBN [44] with multiple layers, and there is a visible and hidden neuron in each of the layer. The visible neurons are fully interconnected with the

hidden neurons. Naturally, the stochastic neuron’s outcome is probabilistic in the Boltzmann networks. The DBN is fully trained to distinguish the occurrence of attackers within the network grounded on the extracted features. DBN framework is an intellectual model that includes of hidden neurons, visible neurons and layers form output layer. Furthermore, there found connotation exists via hidden and input neurons; yet, no relation in visible neurons, the association rule is not existing among hidden neurons. The link existing among visible and hidden neurons is symmetric and exclusive.

The output of the neurons is probabilistic in the Boltzmann network. The output \hat{o} is grounded on the probability function $S(\psi)$ in Eq. (5). The probability function has used the sigmoid-shaped function.

$$\hat{o} = \begin{cases} 1 & \text{with } S(\psi) \\ 0 & \text{with } 1 - S(\psi) \end{cases} \quad (5)$$

$$S(\psi) = \frac{1}{1 + e^{-\frac{\psi}{p}}} \quad (6)$$

Eq. (7) define the DBN model, where, p signifies the pseudo-temperature.

$$\lim_{p \rightarrow 0^+} S(\psi) = \lim_{p \rightarrow 0^+} \frac{1}{1 + e^{-\frac{\psi}{p}}} = \begin{cases} 0 & \text{for } \psi < 0 \\ \frac{1}{2} & \text{for } \psi = 0 \\ 1 & \text{for } \psi > 0 \end{cases} \quad (7)$$

In DBN architecture, the path of the feature processing is shown by a collection of RBM layers, and the classification procedure shown by MLP. The mathematical model depict Boltzmann machine energy in the method of neuron or binary state as portrayed in Eq. (8) and Eq. (9). Where, $w_{c,r}$ indicates the weights amid neurons, which is optimally adjusted or tuned by a new proposed SAEHO model and γ_c specifies the biases.

$$F(\hat{b}) = -\sum_{c < r} \hat{b}_c w_{c,r} - \sum_c \gamma_c \hat{b}_c \quad (8)$$

$$\Delta F(\hat{b}_c) = \sum_r \hat{b}_c w_{c,r} + \gamma_c \quad (9)$$

The growth of energy grounded on combined conformation in visible or hidden neurons (a, f) is described in Eq. (10), Eq. (11) and Eq. (12), where, f_c and a_c portrays the binary state of hidden unit r and c visible unit. X_c and Y_r indicates the biases and $w_{c,r}$ signifies the weight among them.

$$F(\vec{a}, \vec{f}) = -\sum_{(c,r)} w_{c,r} a_c f_r - \sum_c X_c a_c - \sum_r Y_r f_c \quad (10)$$

$$\Delta F(a_c, \vec{f}) = \sum_r w_{c,r} f_r + X_c \quad (11)$$

$$\Delta F(\vec{a}, f_r) = \sum_c w_{c,r} a_c + Y_r \quad (12)$$

RBM training achieves the resultant weight allocation and the distributed probabilities are stated as in Eq. (13). The probability distribution in RBM method for the visible and hidden vectors pair (\vec{a}, \vec{f}) is given in Eq. (14). The partition function Z is specified in Eq. (15).

$$\hat{w}_{(c)} = \max_{\vec{w}} \prod_{\vec{a} \in I} D(\vec{a}) \quad (13)$$

$$D(\vec{a}, \vec{f}) = \frac{1}{Z} e^{-F(\vec{a}, \vec{f})} \quad (14)$$

$$Z = \sum_{\vec{a}, \vec{f}} e^{-F(\vec{a}, \vec{f})} \quad (15)$$

DBN scheme utilize the CD learning approach and the output of DBN is denoted as CL_{DBN} . The final classification output is denoted as CL , and it is expressed in Eq. (16).

$$CL = \frac{CL_{CNN} + CL_{DBN}}{2} \quad (16)$$

VI. PROPOSED SEAGULL ADOPTED ELEPHANT HERDING OPTIMIZATION ALGORITHM FOR OPTIMAL TRAINING OF HYBRID MODEL VIA TUNING THE WEIGHTS

A. Solution Encoding and Fitness Evaluation

The weight of both DBN and CNN are optimally adjusted or tuned via the adopted SAEHO. The input solution subjected to the adopted SAEHO scheme is demonstrated in Fig. 3, where, W_1, W_2, \dots, W_N shows the weights of CNN, w_1, w_2, \dots, w_{mn} shows the weights of DBN, N indicates the total counts of weight in CNN, and mn denotes the total number of weights in DBN. The fitness objective of adopted detection model is stated in Eq. (17). Here, $Loss$ indicates the detection error.

$$Obj = Min(Loss) \quad (17)$$

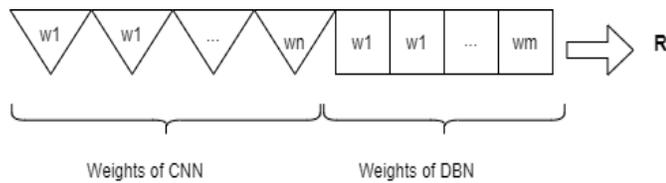


Fig. 3. Solution Encoding.

B. Proposed SAEHO Model

This paper implements a new hybrid SAEHO scheme which combines the logic of EHO [45] and SOA [46], respectively. While the traditional EHO gives better performance; the disadvantage is that it will not use the required data to identify the present and future searches. The existing SOA model has solved the challenging large-scale constrained issues and hence solve 7 constrained real-life industrial applications; still, the constraints are very tedious and computational complexity. This becomes the issue while solving optimization problems. This makes us to combine the logics of both the algorithm since the Hybrid models are reported to be promising for certain search problems with better convergence speed [23]. Here, the logic of SOA is integrated with EHO, and thereby named as SAEHO.

Naturally, the elephants live in social groups called clans and each clan stay with Matriarch, the female elephant leader. The grown male elephant lives separately from their groups. The elephant population is produced randomly and it splits into a number of clans according to their fitness value. EHO algorithm having three major rules to follows.

- The elephant population consists of number of clans with fixed number of female and male elephants in every clan.
- Some of male elephants live far from the clans individually.
- The elephant lives together with the leader of all clans, matriarch (i.e.,) female elephant in each clan.

1) *Clan updating*: In this operator, each clan updating is done individually. Conventionally, the subsequent position is influenced by matriarch h and for each elephant in clan h for the clan updating operators. However, as per the proposed SAEHO method, the SOA updation function is used for clan updation as given in Eq. (18).

$$R_b(\vec{t}) = \hat{B} \times (R_{best}(\vec{t}) - R_t(\vec{t})) + Levy(\zeta) \quad (18)$$

In Eq. (18), R_b indicates the locations of seagull search agent R_t towards the best fit search agent PR_{best} (i.e., fittest seagull), R_t denotes the search agent's current position, \vec{t} refers to the current iteration, and the behavior \hat{B} is randomized used for proper balancing among exploitation and exploration.

Moreover, for the best fit elephant, the updation is achieved by Eq. (19).

$$R_{n,h,k} = \eta \times R_{cen,h} \quad (19)$$

In Eq. (19), $\eta[0,1]$ is the center of clan h . The new individual $R_{n,h,k}$ is expressed from the information obtained by all elephants in clan h . $R_{cen,h}$ represent the centre of clan h , and it is given in Eq. (20).

$$R_{cen,h} = \frac{1}{G_h} \times \sum_{k=1}^{G_h} R_{h,k,\vec{d}} \quad (20)$$

In Eq. (21), $1 \leq \vec{d} \leq \vec{d}$ denotes the \vec{d}^{th} dimension and \vec{d} indicates the total dimensions. G_h specify the number of elephants in clan h . $R_{h,k,\vec{d}}$ refers to the \vec{d} dimensions of the elephant individual $R_{h,k}$.

2) *Separating operator*: The grown male elephants in clan starts live separately. The separating operator is determined after the separating process while solving the optimization problem. As per the proposed SAEHO logic, for enhancing the search ability, the worst fitness of elephant at each generation in the separating operator is defined as per the proposed evaluation given in Eq. (21).

$$R_{worst,h} = R(R_{best} - R_{worst})_{min} \quad (21)$$

Here, R_{min} indicates the minimum bounds in the positions of single elephant. $R_{worst,h}$ indicates the worst elephant individuals of clan ci and $rand$ value is calculated using Chebyshev chaotic map. The value ranges from 0 to 1.

$$R_{\vec{q}+1} = \cos(\vec{q} \cos^{-1}(R_{\vec{q}})) \quad (22)$$

The pseudo code of adopted approach is given in Algorithm 1.

Algorithm 1: Pseudo code of proposed SAEHO method

Initialization

Compute the elephant fitness

Repeat

Arrange all the elephants based on its fitness

Clan updating

For $h = 1$ to n_{clan} do

For $k = 1$ to n_{ci} do

If $R_{h,k} = R_{best,ci}$ then

Clan updation is done by the elephant position using Eq. (20).

Else

Proposed clan updation is done by the seagull position using Eq. (19)

End if

End for k

End h

Separating operator

For $h = 1$ to n_{clan} do

Replace the worst elephant as per the proposed Eq. (22)

End for h

Evaluate population by the new update positions

Until

VI. CONCLUSION

This paper has introduced a novel deep hybrid attack detection method. The input data subjected for preprocessing phase and data normalization process was carried out. From the preprocessed data, the statistical and higher order statistical features were extracted. Finally, the extracted features were given to hybrid deep learning model for detecting the presence of attack. The proposed hybrid classifier combines the models like DBN and CNN. To make the detection more precise and accurate algorithm named SAEHO proposed which used for tuning the optimal weights. SAEHO combines the logic of EHO and SOA. In the algorithm, two kinds of operator used i.e., clan updating and separating. Clan updating is done by EHO providing it able to find the best position else SOA is used to find the solution. As our next task, the performance of the proposed model will be computed over the present methods in terms of various metrics like FNR, MCC, Rand index, sensitivity, FPR, specificity, FDR, precision, NPV, accuracy, and FMS, correspondingly.

REFERENCES

- [1] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network Intrusion Detection for IoT Security Based on Learning Techniques," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 3, pp. 2671–2701, Jul. 2019, doi: 10.1109/COMST.2019.2896380.
- [2] S. Rathore and J. H. Park, "Semi-supervised learning based distributed attack detection framework for IoT," *Applied Soft Computing Journal*, vol. 72, pp. 79–89, Nov. 2018, doi: 10.1016/j.asoc.2018.05.049.
- [3] M. Hossain and J. Xie, "Third Eye: Context-Aware Detection for Hidden Terminal Emulation Attacks in Cognitive Radio-Enabled IoT Networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 1, pp. 214–228, Mar. 2020, doi: 10.1109/TCCN.2020.2968324.
- [4] S. M. Tahsien, H. Karimipour, and P. Spachos, "Machine learning based solutions for security of Internet of Things (IoT): A survey," *Journal of Network and Computer Applications*, vol. 161, Jul. 2020, doi: 10.1016/j.jnca.2020.102630.
- [5] A. Kore and S. Patil, "IC-MADS: IoT Enabled Cross Layer Man-in-Middle Attack Detection System for Smart Healthcare Application," *Wireless Personal Communications*, vol. 113, no. 2, pp. 727–746, Jul. 2020, doi: 10.1007/s11277-020-07250-0.
- [6] D. Yin, L. Zhang, and K. Yang, "A DDoS Attack Detection and Mitigation with Software-Defined Internet of Things Framework," *IEEE Access*, vol. 6, pp. 24694–24705, Apr. 2018, doi: 10.1109/ACCESS.2018.2831284.
- [7] A. Y. Khan, R. Latif, S. Latif, S. Tahir, G. Batool, and T. Saba, "Malicious Insider Attack Detection in IoTs Using Data Analytics," *IEEE Access*, vol. 8, pp. 11743–11753, 2020, doi: 10.1109/ACCESS.2019.2959047.
- [8] K. Mandal, M. Rajkumar, P. Ezhumalai, D. Jayakumar, and R. Yuvarani, "Improved security using machine learning for IoT intrusion detection system," *Materials Today: Proceedings*, Dec. 2020, doi: 10.1016/j.matpr.2020.10.187.
- [9] D. C. Wang, I. R. Chen, and H. Al-Hamadi, "Reliability of Autonomous Internet of Things Systems with Intrusion Detection Attack-Defense Game Design," *IEEE Transactions on Reliability*, vol. 70, no. 1, pp. 188–199, Mar. 2021, doi: 10.1109/TR.2020.2983610.
- [10] S. Rani and N. Singh Gill, "HYBRID MODEL FOR TWITTER DATA SENTIMENT ANALYSIS BASED ON ENSEMBLE OF DICTIONARY BASED CLASSIFIER AND STACKED MACHINE LEARNING CLASSIFIERS-SVM, KNN AND C5.0," *Journal of Theoretical and Applied Information Technology*, vol. 29, p. 4, 2020, Accessed: Jan. 19, 2022. [Online]. Available: www.jatit.org
- [11] N. S. G. Sangeeta Rani, "Hybrid Model using Stack-Based Ensemble Classifier and Dictionary Classifier to Improve Classification Accuracy of Twitter Sentiment Analysis," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 7, 2020.
- [12] P. Gulia and N. Singh Gill, "Comprehensive Analysis of Flow Incorporated Neural Network based Lightweight Video Compression Architecture," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 12, no. 3, 2021, Accessed: Jan. 19, 2022. [Online]. Available: www.ijacsa.thesai.org
- [13] Y. Jia, F. Zhong, A. Alrawais, B. Gong, and X. Cheng, "FlowGuard: An Intelligent Edge Defense Mechanism against IoT DDoS Attacks," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9552–9562, Oct. 2020, doi: 10.1109/JIOT.2020.2993782.
- [14] T. Zhi, Y. Liu, and J. Wu, "A Reputation Value-Based Early Detection Mechanism against the Consumer-Provider Collusive Attack in Information-Centric IoT," *IEEE Access*, vol. 8, pp. 38262–38275, 2020, doi: 10.1109/ACCESS.2020.2976141.
- [15] H. Al-Hamadi, I. R. Chen, D. C. Wang, and M. Almashan, "Attack and defense strategies for intrusion detection in autonomous distributed IoT systems," *IEEE Access*, vol. 8, pp. 168994–169009, 2020, doi: 10.1109/ACCESS.2020.3023616.
- [16] S. Patranabis et al., "Lightweight Design-for-Security Strategies for Combined Countermeasures Against Side Channel and Fault Analysis in IoT Applications," *Journal of Hardware and Systems Security*, vol. 3, no. 2, pp. 103–131, Jun. 2019, doi: 10.1007/s41635-018-0049-y.
- [17] A. Sagu, N. Singh, G., and P. Gulia, "Artificial Neural Network for the Internet of Things Security," *International Journal of Engineering Trends and Technology*, vol. 68, pp. 137–144, 2020, doi: 10.14445/22315381/IJETT-V68I1P218.
- [18] N. S. G. Sagu Amit, "Machine Learning Decision Tree Classifier and Logistics Regression Model," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 1.4, pp. 163–166, Sep. 2020, doi: 10.30534/ijatcse/2020/2491.42020.
- [19] N. S. G. Sangeeta, "Framework for Tweet Sentiment Classification Using Boostingbased Ensemble Approach," *CIENCIA E TECNICA.VITIVINICOLAA SCIENCE AND TECHNOLOGY JOURNAL (ISSN: 2416-3953)*, pp. 1–13, 2017.
- [20] S. Rani, N. S. Gill, and P. Gulia, "Analyzing impact of number of features on efficiency of hybrid model of lexicon and stack based ensemble classifier for twitter sentiment analysis using WEKA tool," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, pp. 1041–1051, May 2021, doi: 10.11591/IJEECS.V22.I2.PP1041-1051.
- [21] P. Gulia, "Performance Analysis of Advancements in Video Compression with Deep Learning," *International Journal of Electrical Engineering and Technology*, vol. 11, no. 5, pp. 137–143, 2020, doi: 10.34218/IJEET.11.5.2020.016.

- [22] M. Zaminkar and R. Fotuhi, "SoS-RPL: Securing Internet of Things Against Sinkhole Attack Using RPL Protocol-Based Node Rating and Ranking Mechanism," *Wireless Personal Communications*, vol. 114, no. 2, pp. 1287–1312, Sep. 2020, doi: 10.1007/s11277-020-07421-z.
- [23] M. M. Beno, V. I. R, S. S. M, and B. R. Rajakumar, "Threshold prediction for segmenting tumour from brain MRI scans," *International Journal of Imaging Systems and Technology*, vol. 24, no. 2, pp. 129–137, 2014, doi: 10.1002/ima.22087.
- [24] R. Marimuthu and B. Chakraborty, "An Approach for Speech Enhancement Using Deep Convolutional Neural Network," 2019.
- [25] A. Sarkar, "Optimization Assisted Convolutional Neural Network for Facial Emotion Recognition."
- [26] Ganeshan R, "Skin Cancer Detection with Optimized Neural Network via Hybrid Algorithm."
- [27] Vinolin V and Vinusha S, "Resbee Publishers Journal of Computational Mechanics, Power System and Control Enhancement in Biodiesel Blend with the Aid of Neural Network and SAPSO," 2018.
- [28] J. Bhasha Shaik, "Resbee Publishers Journal of Computational Mechanics, Power System and Control Deep Neural Network and Social Ski-Driver Optimization Algorithm for Power System Restoration with VSC-HVDC Technology."
- [29] Bhagyalakshmi V, DrRamchandra, and DrGeeta D, "Resbee Publishers Journal of Networking and Communication Systems Arrhythmia Classification Using Cat Swarm Optimization Based Support Vector Neural Network."
- [30] S. B. Chandanapalli, S. Reddy, and R. Lakshmi, "Resbee Publishers Journal of Networking and Communication Systems Convolutional Neural Network for Water Quality Prediction in WSN," 2019.
- [31] N. Ravi and S. M. Shalinie, "Learning-Driven Detection and Mitigation of DDoS Attack in IoT via SDN-Cloud Architecture," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3559–3570, Apr. 2020, doi: 10.1109/JIOT.2020.2973176.
- [32] J. Yoon, "Deep-learning approach to attack handling of IoT devices using IoT-enabled network services," *Internet of Things (Netherlands)*, vol. 11, Sep. 2020, doi: 10.1016/j.iot.2020.100241.
- [33] J. Bhayo, S. Hameed, and S. A. Shah, "An Efficient Counter-Based DDoS Attack Detection Framework Leveraging Software Defined IoT (SD-IoT)," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.3043082.
- [34] D. Li, Z. Cai, L. Deng, X. Yao, and H. H. Wang, "Information security model of block chain based on intrusion sensing in the IoT environment," *Cluster Computing*, vol. 22, pp. 451–468, Jan. 2019, doi: 10.1007/s10586-018-2516-1.
- [35] J. Roldán, J. Boubeta-Puig, J. Luis Martínez, and G. Ortiz, "Integrating complex event processing and machine learning: An intelligent architecture for detecting IoT security attacks," *Expert Systems with Applications*, vol. 149, Jul. 2020, doi: 10.1016/j.eswa.2020.113251.
- [36] M. V. O. de Assis, L. F. Carvalho, J. J. P. C. Rodrigues, J. Lloret, and M. L. Proença, "Near real-time security system applied to SDN environments in IoT networks using convolutional neural network," *Computers and Electrical Engineering*, vol. 86, Sep. 2020, doi: 10.1016/j.compeleceng.2020.106738.
- [37] K. Mabodi, M. Yusefi, S. Zandiyan, L. Irankhah, and R. Fotuhi, "Multi-level trust-based intelligence schema for securing of internet of things (IoT) against security threats using cryptographic authentication," *Journal of Supercomputing*, vol. 76, no. 9, pp. 7081–7106, Sep. 2020, doi: 10.1007/s11227-019-03137-5.
- [38] S. Huang, C. Lin, K. Zhou, Y. Yao, H. Lu, and F. Zhu, "Identifying physical-layer attacks for IoT security: An automatic modulation classification approach using multi-module fusion neural network," *Physical Communication*, vol. 43, Dec. 2020, doi: 10.1016/j.phycom.2020.101180.
- [39] H. Mohammed, S. R. Hasan, and F. Awwad, "Fusion-on-field security and privacy preservation for IoT edge devices: Concurrent defense against multiple types of hardware trojan attacks," *IEEE Access*, vol. 8, pp. 36847–36862, 2020, doi: 10.1109/ACCESS.2020.2975016.
- [40] R. Sahay, G. Geethakumari, and B. Mitra, "A novel blockchain based framework to secure IoT-LLNs against routing attacks," *Computing*, vol. 102, no. 11, pp. 2445–2470, Nov. 2020, doi: 10.1007/s00607-020-00823-8.
- [41] S. A. Alabady, F. Al-Turjman, and S. Din, "A Novel Security Model for Cooperative Virtual Networks in the IoT Era," *International Journal of Parallel Programming*, vol. 48, no. 2, pp. 280–295, Apr. 2020, doi: 10.1007/s10766-018-0580-z.
- [42] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier," *Expert Systems with Applications*, vol. 36, no. 7, pp. 10618–10626, Sep. 2009, doi: 10.1016/j.eswa.2009.02.053.
- [43] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *ISCAS 2010 - 2010 IEEE International Symposium on Circuits and Systems: Nano-Bio Circuit Fabrics and Systems*, 2010, pp. 253–256. doi: 10.1109/ISCAS.2010.5537907.
- [44] H. Z. Wang, G. B. Wang, G. Q. Li, J. C. Peng, and Y. T. Liu, "Deep belief network based deterministic and probabilistic wind speed forecasting approach," *Applied Energy*, vol. 182, pp. 80–93, Nov. 2016, doi: 10.1016/j.apenergy.2016.08.108.
- [45] M. A. Elhosseini, R. A. el Sehiemy, Y. I. Rashwan, and X. Z. Gao, "On the performance improvement of elephant herding optimization algorithm," *Knowledge-Based Systems*, vol. 166, pp. 58–70, Feb. 2019, doi: 10.1016/j.knosys.2018.12.012.
- [46] G. Dhiman and V. Kumar, "Seagull optimization algorithm: Theory and its applications for large-scale industrial engineering problems," *Knowledge-Based Systems*, vol. 165, pp. 169–196, Feb. 2019, doi: 10.1016/j.knosys.2018.11.024.

A Node Monitoring Agent based Handover Mechanism for Effective Communication in Cloud-Assisted MANETs in 5G

B.V.S Uma Prathyusha¹

Research Scholar
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore, India

K.Ramesh Babu^{2*}

Professor
School of Computer Science and Engineering
Vellore Institute of Technology, Vellore, India

Abstract—As nodes often join or leave the network, the communication between the cloud and the MANET remains unreliable in Cloud-Assisted MANET. The event of connection failure in MANET presents several challenges to the network, in particular, the handover issue and high energy consumption during route re-establishment if a connection fails in D2D (device-to-device) communication networks. To address this problem of D2D mobile communication in 5G, we propose a Node Monitoring Agent Based Handover Mechanism (NMABHM). To improve the network's efficiency, we use the K-means algorithm for clustering and cluster head selection in Hybrid MANET and maintain a backup routing table based on a Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) to quickly recover the route. Additionally, a Node Monitoring Agent (NMA) is introduced to handle the handover issue if a node comes out of range during the communication phase. The NMABHM-based handover mechanism is proposed with group mobility over a cluster-based architecture involving agents. The results of the simulation indicate that, in terms of lower energy consumption and higher throughput, our proposed mechanism is more effective than the current routing mechanisms.

Keywords—MANET; clustering; cloud computing; 5G Wireless networks; cloud-assisted MANET

I. INTRODUCTION

Wireless technology, the most important communication medium for information between different devices, has witnessed rapid growth over the subsequent years. The future wireless technology, i.e. 5G, which is a new global standard of communication, is especially designed for the purpose of forward compatibility and to flexibly support services like mission-critical communication, mobile broadband, IOT, cloud-based services, and D2D communication [1]. With the increasing demand for QoS and the data transfer speed for mobile networks, there is still a requirement for new communication technologies [2,3]. Due to the compatibility features, Cloud Assisted MANET, which is the combination of Cloud and Mobile Ad-Hoc Network, has received a lot of interest from researchers today.

A. MANET

MANET is one of the most widely used Ad-Hoc wireless networks [4]. Routing in MANET is multi-hop and each node

in the network is a transceiver with bidirectional links. MANET's distinguishing feature, in addition to its unique features and advantages like dynamic topology, mobility, multi-hop, and decentralised administration, is its ease of connection with other networks. With this superiority in communication, MANET has made significant contributions to the growth of the Internet industry [3].

B. Cloud Computing

Cloud Computing has been observed exponential growth in recent developments and the potential has been developed to limit energy consumption and to enhance the usefulness of mobile nodes in MANET [5]. Cloud storage platforms are used by the three administrative models, like IaaS (Infrastructure as a Service), SaaS (Software as a Service), and PaaS (Platform as a Service), which are delivered to users through Internet-connected systems [6]. In addition to the benefits of Cloud Computing like scalability, affordability, and security, the 5G and D2D connectivity specifications require the use of CA-MANETs, resulting in various research gaps in this area.

C. CA-MANET

The convergence of MANET and Cloud Computing has become more popular in recent years. MANETs are widely used in emergency situations because they don't adhere to a predetermined topology. When disasters like earthquakes or other natural calamities strike, smart devices in MANET will need to access the data on health care services and other essentials stored on site maps via the Cloud. The requirement for mobile devices in MANET to access cloud services, as well as D2D communication and 5G wireless networks, necessitated the CA-MANET architecture, where multiple devices may be simply connected. CA-MANET unquestionably constructs robust 5G networks [7]. An overlay of peers, i.e. mobile devices and cloud data servers, is formed in CA-MANET, as shown in Fig. 1, when the cloud server joins the MANET [8,9]. The overlay in CA-MANET is self-organizing, i.e. any Peer can join or leave the network anytime. In CA-MANET, Super Peers are the cloud data servers interconnected logically to the Peers. In order to route, search, and send messages, each Peer keeps the information necessary to do so and may be linked to the Super Peer either directly or indirectly. If a Peer and a Super Peer have an indirect link, intermediate Peers operate as routers and help forward the requested services.

*Corresponding Author.

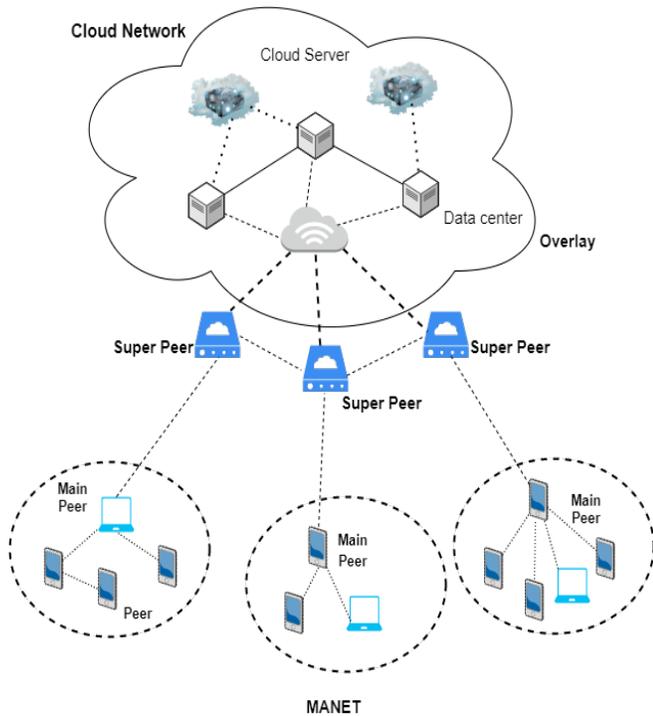


Fig. 1. An Overlay of Data Centers and Cloud Servers with D2D Communication.

The role of cloud servers in the overlay is to provide the required data and services for MANET communication. When MANET requests a particular cloud service, the cloud server performs operations like searching, routing, and updating its database in order to provide future service requests [10]. Due to frequent link breakages in MANET, there is a chance for communication loss between the cloud and MANET. This creates several of the issues like handover, link failures, an increase in overhead and energy consumption in route re-discovery, which indeed effects the cloud servers' service delivery efficiency [11,12]. The cost of processing power and energy consumption grows as network resources are depleted, affecting D2D network performance in 5G networks, which is the main challenge. In addition to that, link failures, effective handover, bandwidth utilization, and route selection are also other challenges to consider. Analyzing the above concerns in detail indicates that energy consumption is the key concern, which is dependent on MANET routing protocols [31]. If energy degradation is handled by selecting the most robust and optimal MANET routing approach, most CA-MANET problems will be solved. For this, we propose a Node Monitoring Agent-Based Handover Mechanism (NMABHM) that focuses on the minimization of energy consumption by effectively handling link failure and service handover.

The significant contributions of the proposed work include:

- Organizing the MANET into Clusters and electing the Cluster Head (CH) based on the K-means algorithm that benefit in bandwidth reusability.

- Introducing a Node Monitoring Agent (NMA) that can perform the ease of handover process without failure by supervising the link failures and node movement.
- Effective Route establishment and maintenance by introducing TOPSIS [13,14] procedure in the AODV algorithm that benefits in maintaining backup route table for selecting optimal alternate paths during link failure.

The rest of the paper is structured in the following manner: Section 2 describes existing works related to minimizing energy consumption in CA-MANET. Section 3 describes the problem description and scenario. Section 4 describes our proposed NMABHM. Section 5 provides performance evaluation and results, followed by Section 6 of Conclusion and Future Work.

II. LITERATURE SURVEY

Due to the growing popularity and increasing demand of mobile devices in wireless communication, 5G networks have evolved. Because of the limited resources available, 5G network components must be able to efficiently use energy and power [15]. To increase overall network performance, it is necessary to study the energy consumption of network nodes [16]. This can be accomplished by the deployment of intelligent approaches to make optimal routing decisions in communication.

In [17,18] the authors suggested a new protocol with a sophisticated costing function that uses residual battery capacity and hop count as parameters to determine high throughput, low energy and long-lasting data transmission routes on the basis of mobile agent technology. In [19] the authors proposed a new strategy called SBR-Stable Backup Routing, which includes techniques for the design and management of backup routes. The SBR greatly reduces packet loss and increases delivery ratio by overhearing MAC signals and the bit error rate of networks. In order to improve the efficiency of the OLSR routing protocol, the authors in [20] developed a method, termed AIS-OLSR, that uses an artificial immune system. Negative selection and ClonalG algorithms are used to calculate hop count, residual energy in relay nodes, and distance between nodes.

In [21] the authors presented ELBRP, which focuses on the problem of energy-awareness by analysing the nodes' amounts of energy and various forwarding strategies that decrease the energy utilization, optimize the delay, and enhance the network usage. In the survey on mobile device energy usage [22], the authors concentrated on various offloading mechanisms that are responsible for the low battery life of mobile phones due to their unique components. In [23] the authors conducted a detailed assessment of the MANET network challenges and discussed the prospect of resolving them using neural network-based clustering by using a multi-criteria decision of network characteristics, specific cluster algorithm implementations for routing applications. To optimize energy consumption of wireless sensor networks, the authors in [24] suggested an effective cluster head selection approach utilising the K-means algorithm. In this method, the cluster head is identified by

reducing the sum of Euclidean distances amongst the head and its members.

In [8] the authors proposed a novel collaboration paradigm between cloudlets and MANETs, in which the objective is to conserve energy and benefit from green computing by developing DCRM. In [9,18] the authors proposed an overlay architecture, in which a viable energy saving strategy is used to reduce searching and routing activities in cloud data servers.

Due to the high degree of node mobility in MANET, handover happens when a node in communication moves out of the coverage ratio. The functionality of the handover requires a time delay. The difficult part is to identify the appropriate handover and execute it well. In [29] the authors proposed a Vertical handover using MIH/SDN to optimise handover in the future generation of mobile networks. In [25] the authors proposed the IMMh mobility handover strategy for IPv6-based MANETs. In IMMh, a user always communicates with a node via its home address, and both the node's mobility handover and care-of address updating processes eliminate the cost and latency associated with duplicate address detection. In [26] the authors address the handover problem for a fixed path by introducing a cloud-assisted ant colony-based solution termed CAFp. This solution leverages Cloud Computing to significantly reduce the time required for handover, notably during handover judgment.

The energy issue is among the most difficult issues to deal with when using cloud networks and MANETs together. When there is a disconnect between a MANET and the cloud, it might result in a cost issue. Many studies have offered routing options for MANETs that save energy and increase performance. EERR is an energy-efficient and effective routing method that has been developed by the authors in [27] for mobility prediction-based MANET systems. EERR uses location information to determine the best transmission power for delivering packets successfully. In [10,18] the authors

proposed CEPRM method, in which service requests can be directed directly to the destination cloud server by preserving a content map on the Super Peer nodes. It decreases the cost of searching time and so minimises the amount of energy consumed by the cloud server's search operation. To help CA-MANETs recover quickly when a connection fails, the authors in [28] developed an EECRM -Energy Efficient Cloud-Assisted Routing Mechanism. In this mechanism, the Bellman-Ford algorithm is altered with parameters such as residual and total energy of the nodes in the network. Backup nodes discovered via nearby nodes during the route maintenance stage aid quick route restoration by presenting alternate paths that are discovered, which reduces energy exhaustion. According to existing research, it is clear that high energy consumption and connection handover issues in CA-MANET's are still unsolved [30]. Also, most of the currently available techniques, such as CEPRM and EEPRM, are proposed for optimal route recovery and maintenance in the event of a link failure between the Main Peer and the Super Peer. However, none of these approaches took into account the scenario of establishing and maintaining an effective route for the Indirect Peer when it moves out of range during the on-going communication process. This contributes to the motivation for the proposed NMABHM, which successfully establishes a path between the Indirect Peer and the Main Peer, by utilising residual energy and handover procedures effectively.

III. PROBLEM DESCRIPTION

Fig. 2 describes the problem scenario in CA-MANET that has been considered for the proposed work. Here, Peer A is the Main Peer as it connects directly to the Super Peer S1. The services provided by Super Peer S1 can be accessed by the Indirect Peers via Main Peer A. Now the Indirect Peer C requests service from S1 through A. If the service requested by Peer C is in the cache of S1, then S1 can respond to the service through Main Peer A.

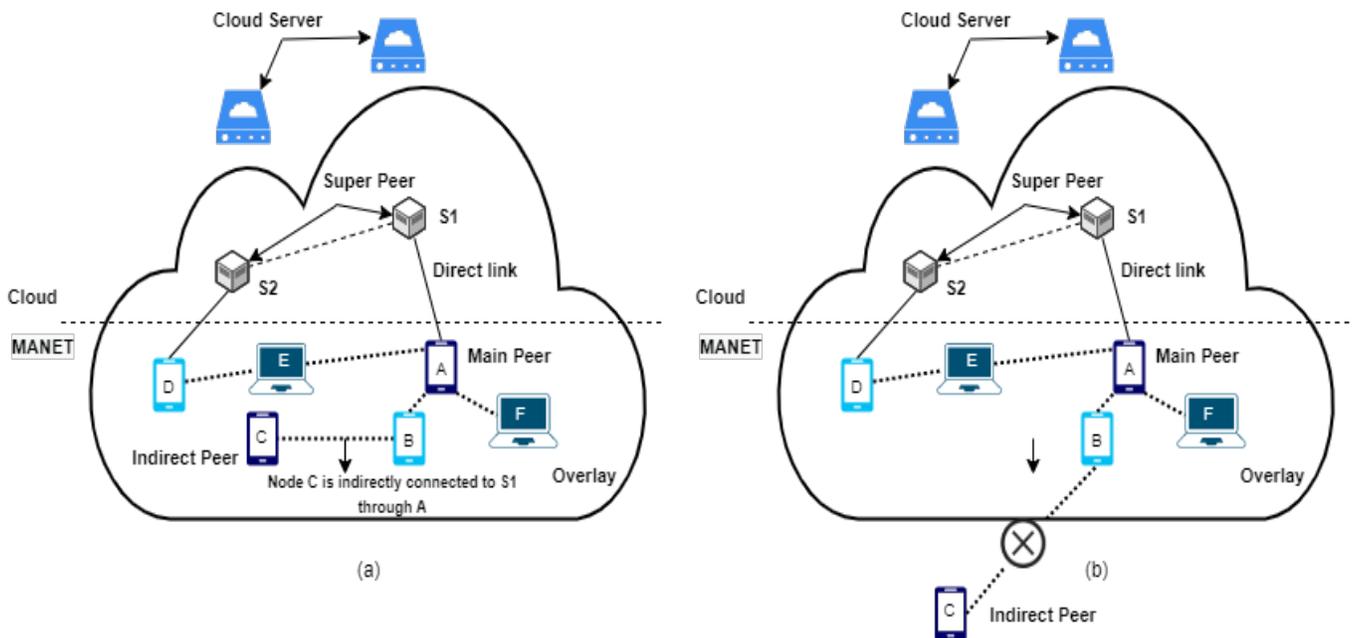


Fig. 2. Link Failure in CA-MANET during Ongoing Communication Process.

Assume that the requested service is in the cache of S1 and the communication is initiated between the Super Peer S1 and the Indirect Peer C through Main Peer A, as shown in Fig. 2(a). During the ongoing communication process between Super Peer S1 and indirect Peer C through Main Peer A and relay Peer B, imagine that Indirect Peer C begins to move away from Main Peer A. As shown in Fig. 2(b), when the indirect Peer C comes out of range from Main Peer A, a link failure occurs, resulting in communication loss between Super Peer S1 and the indirect Peer C. This interrupts the ongoing service communication between Super Peer S1 and Indirect Peer C, as the Indirect Peer C is now out of range from the Main Peer A to access the service. This is due to loss of connectivity and it results in a problem of incomplete service requests. This problem can be resolved by the effective handover of the requested service when the Indirect Peer moves out of range of its service resource.

A. Approach

To address this problem, we propose a Node Monitoring Agent-Based Handover Mechanism (NMABHM), in which the main idea is to implement a mechanism that allows the Indirect Peer to have access to the requested service from the cloud, in the case when it moves out of range from the Super Peer. For the proposed work, we follow an approach based on the following assumptions in CA-MANET:

- A Hybrid MANET scenario is implemented using a cluster-based architecture and the Cluster Head (CH) is considered the Main Peer.
- CH maintains a direct link to the Super Peer to communicate with the cloud services.
- CHs are all static in nature.
- Using the CEPRM algorithm, the Super Peers can share service information with one another and with additional Super Peers when they enter an overlay [10].

IV. PROPOSED MECHANISM

A. NMABHM Algorithm

Fig. 3 represents the CA-MANET considered for the proposed work, in which the NAMBHM operates in three phases as follows:

1) *Step 1*: Using a K-means algorithm to group the mobile nodes for reliable and successful data transmission and Optimal Cluster Head Selection by following K-means algorithm based on the Euclidean distance and residual energy.

2) *Step 2*: Implementing a Node Monitoring Agent (NMA) to perform effective handover in the event of a link failure.

3) *Step 3*: Introducing the TOPSIS technique into the AODV algorithm for Effective Route establishment and Maintenance.

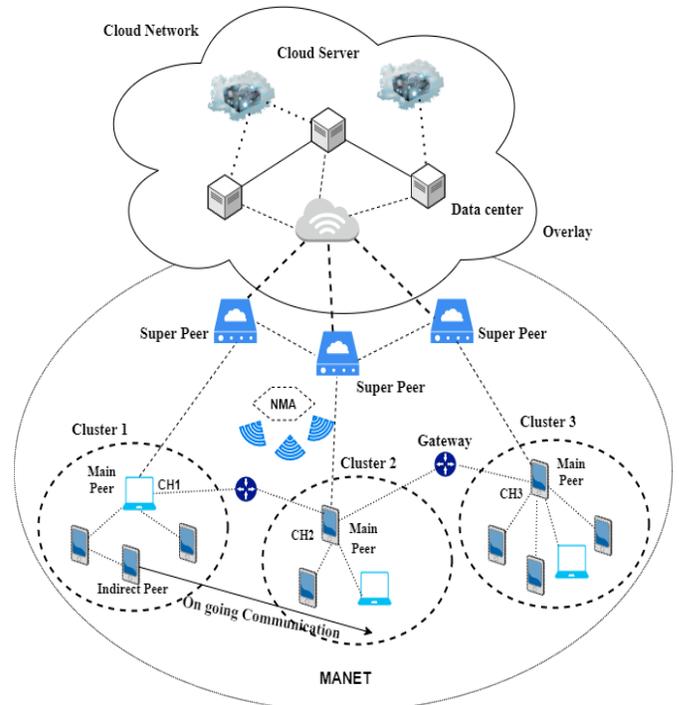


Fig. 3. Indirect Peer Movement from Cluster 1 to Cluster 2 during Ongoing Communication Process.

B. Clustering and CH Selection using K-Means Algorithm

1) *Clustering*: Clustering is still the most extensively utilised performance optimization technique for MANETs, enabling them to expand to a large number of mobile nodes. In order to perform the clustering in the network, a K-means algorithm is applied to group the mobile nodes, which creates clusters of nodes based on their Euclidean distances from one another [28]. This technique is used to generate clusters, so that the distance between nodes and the CH is as short as possible. For this, we suppose that most nodes have understood their particular location information by using either a GPS unit or a network localization technique.

2) *Cluster head selection*: The CH selection procedure is divided into three steps, which are as follows:

a) *Step 1*: The First step is initial clustering.

The K-means technique is used to form clusters among the nodes within the network [24]. Assume that the n-node network is partitioned into k-clusters. Then, k nodes out of n are chosen at random to be the CHs. Based on the Euclidean distance, each of the available node chooses the CH nearest to it.

b) *Step 2*: The next step is to re-cluster.

The centroid of each cluster with an “s” number of nodes is computed after assigning each node to one of the network’s k-clusters.

$$\text{Centroid}(X, Y) = \left(\frac{1}{s} \sum_{i=1}^s x_i, \frac{1}{s} \sum_{i=1}^s y_i \right) \quad (1)$$

c) *Step 3*: The third step is to choose a Cluster Head.

After generating the clusters, each node is assigned an ID number in the cluster based on its centroid distance, such that the closer nodes receive smallest ID number. A node's ID number determines the sequence in which it shall be elected as the Cluster Head. As a result, the ID number is critical in determining which node is designated as CH.

In each round, the CHs residual energy is verified to maintain the networks connectivity link. If the energy of the CH is less than the predefined threshold, the next order node is chosen as the new CH. The newly elected CH communicates the change in the CH to other nodes. This approach employs a routing mechanism that enables CHs to communicate directly with the Super Peer through a single-hop connection. The Super Peer then processes the collected information.

C. Node Monitoring Agent based Handover Mechanism

Once the Cluster Heads (CHs) are selected, Node Monitoring Agent Based Handover Mechanism (NMABHM) is incorporated with group mobility over cluster-based architecture involving agents in the CA-MANET for effective handover of service during the link breakage. This mechanism comprises all the correspondent properties for making handover decision. The various properties considered by the NMA are described as follows:

1) *Capacity of cluster*: This property ensures that the number of nodes available in a cluster is always lower than the original capacity of nodes that could be serviced by the CH in the corresponding cluster. Failure of this property results in handover.

2) *Cloud Services*: If the cloud service demanded by the mobile host is the same as that supplied by CH, the Peer must be present in the cluster, otherwise, according to the application specifications, it must scan for another network.

3) *Received signal strength indication (RSSI)*: If the mobile host has same RSSI values from two different CHs it is then important to search the next near-hop and decide which cluster it will have to communicate with.

4) *NMA*: NMA is a handover issue detection agent that supervises various network events and updates the collection of CHs and is ultimately responsible for property-based handover decisions. It is responsible for interacting with CHs and Super Peers with various parameter-based decisions to choose the appropriate policy and providing CHs with suggestions. If anyone of the following properties fails, then handover occurs.

5) *Property 1*: Capacity of cluster

If the nodes count in a cluster is greater than the capacity, then

Set value for the parameter as:

pr_{var1} : No. of nodes to make handover

pr_{var2} : D is distance in terms of hop count to new cluster head.

pr_{var3} : Handover latency (H_{lay})

6) *Property 2*: Cloud Services

When looking for specified services on mobile hosts, the parameter values are

pr_{var1} : Offered Service list

pr_{var2} : Offered service plans

pr_{var3} : Security features provided

7) *Property 3*: Next Hop Specification

If two mobile hosts receive same RSSI from two or more different cluster heads, then

Set parameter Value as:

pr_{var1} : Min $P_{Consumed}$ (MN)

pr_{var2} : RSSI value

pr_{var3} : Weight variable value

To perform these steps, dynamic packets are created. Those packets are known as Request Packets of Handover (RPH). Then handoff is predicted via Handoff Label Packet (HLP) as shown in Table I.

TABLE I. PACKET INFORMATION

Request Packet of Handover (RPH)		
pr_{var1}	pr_{var2}	pr_{var3}
Handoff Label Packet (HLP)		
RSSI value	H_{lay}	Min $P_{Consumed}$ (MN)

D. Optimal Route Selection using TOPSIS Algorithm

In the proposed work, nodes can communicate with each other by using the AODV routing protocol and a backup routing table is maintained based on two parameters such as the hop count and the available energy. In case of an ongoing communication process, when the communicating Indirect Peer moves out of range from the Main Peer (CH), communication loss occurs between the Main Peer and the communicating Indirect Peer which results in a connectivity issue between the Super Peer and the communicating Indirect Peer. Now, to utilize the requested service, the Indirect Peer must be connected to the Main Peer (CH) of that particular region in which the Indirect Peer is present i.e. a route must be established between the Cluster Head and the Indirect Peer.

1) Route establishment and maintenance:

- The node movement information is shared by the NMA with all the Cluster Heads in the CA-MANET.
- On receiving the node movement information, the CH broadcasts the HELLO messages to all of the neighbouring nodes in the cluster.
- The Indirect Peer broadcasts RREQ and route establishment is done between the CH (Main Peer) and the Indirect Peer as shown in Fig. 4.

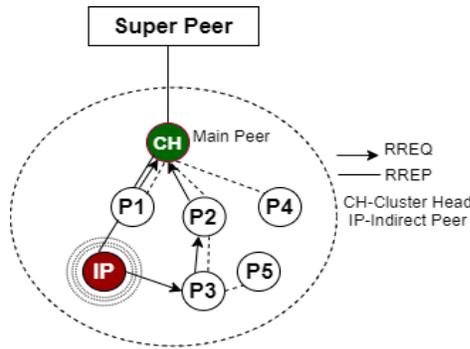


Fig. 4. Route Establishment and Maintenance when Indirect Peer Enters into Another Cluster.

2) TOPSIS process:

The following measures include the TOPSIS process:

a) Step 1: To define the Decision Matrix (DM)

$$DM = \begin{matrix} & T_1 & T_2 & \dots & T_n \\ \begin{matrix} AP_1 \\ AP_2 \\ \vdots \\ AP_m \end{matrix} & \begin{bmatrix} t_{11} & t_{12} & \dots & t_{1n} \\ t_{21} & t_{22} & \dots & t_{2n} \\ \vdots & \vdots & \dots & \vdots \\ t_{m1} & t_{m2} & \dots & t_{mn} \end{bmatrix} \end{matrix} \quad (2)$$

In this case, $i = 1..m$ represents the criterion index, m represents the number of possible connections, and $j = 1..n$ represents the alternative index [13,14]. The criteria are T_1, T_2, \dots, T_n , and the alternate positions are AP_1, AP_2, \dots, AP_m . The matrix components in alternative j are linked to the set of parameters i .

b) Step 2: Create a Normalized Decision Matrix (NDM) that accurately represents the output of design alternatives.

$$NDM = R_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^m x_{ij}^2}} \quad (3)$$

c) Step 3: Evaluate the weight decision Matrix by considering the random weight as W_j .

$$V = V_{ij} = W_j \times R_{ij} \quad (4)$$

d) Step 4: Using the weighted decision matrix the Ideal solutions are generated.

$$I^+ = \{v_1^+, v_2^+, v_3^+, \dots, v_n^+\} \quad (5)$$

in which $V_j^+ = \{(max_i(v_{ij}) \text{ if } j \in J); (min_i v_{ij} \text{ if } j \in J')\}$

$$I^- = \{v_1^-, v_2^-, v_3^-, \dots, v_n^-\} \quad (6)$$

in which $V_j^- = \{(min_i(v_{ij}) \text{ if } j \in J); (max_i v_{ij} \text{ if } j \in J')\}$

where, I^+ is the Positive Ideal Solution and I^- is the Negative Ideal Solution [14]. The beneficial attributes are given with J and non-beneficial attributes with J' .

e) Step 5: Using the above Ideal solution i.e., I^+ and I^- , compute each competitive alternative's separation distance (SD) with the criteria index 'i' and the alternative index 'j'.

$$SD^+ = \sqrt{\sum_{j=1}^n (V_j^+ - V_{ij})^2} \text{ for } i = 1, \dots, m \quad (7)$$

$$SD^- = \sqrt{\sum_{j=1}^n (V_j^- - V_{ij})^2} \text{ for } i = 1, \dots, m \quad (8)$$

f) Step 6: Calculate the optimal solution for all the locations depending on their proximity.

The relative closeness (RC) of each possible position within the optimal solution is determined for each competitive alternative.

$$RC_i = \frac{SD_i^-}{(SD_i^+ + SD_i^-)}, 0 \leq RC_i \leq 1 \quad (9)$$

g) Step 7: Prioritize the alternatives in order of preference.

The greater the value for relative proximity, the better the order of rating, and thus, dependent on the value of C_i , the greater the alternative output. In descending order, the rating of preference thus makes it possible to equate comparatively better results.

V. PERFORMANCE EVALUATION

A. Simulation Setup

To evaluate the simulation findings with other research work, the standard simulation environment for the NS2 simulator has been implemented using this mechanism. Table II presents the simulation parameters considered for the network.

B. Result and Analysis

By assuring effective handover, the proposed NMABHM addresses the issue of energy consumption in CA-MANETs. For this purpose, the execution process period and energy consumption are chosen as the parameters for performance evaluation. Energy consumption is calculated based on execution time, and a comparison is made between our Node Monitoring Agent-based Handover Mechanism-(NMABHM) and the existing Energy-Efficient Cloud-Assisted Routing Mechanism-(EECRM).

TABLE II. SIMULATION PARAMETERS: NS2 SIMULATOR

Parameter	Value
Area	1000*500
No of nodes	40
Queue Length	50
Protocol	NMABHM (modifying AODV)
Model	Energy Model
Initial Energy	100 J
Topology	Flat Grid
Simulation time	400 sec

The other parameters considered are:

1) *Distance between node measurements in NS2:* A Dynamic topology (coordinates x and y) is created. The distance between the nodes is calculated by applying the coordinates of the node (x, y) in the Pythagorean theorem.

2) *Available energy in NS2:* The level of energy in the network is described by utilizing the energy model. In this energy model, the initial value, called "initial energy," is the energy level that the node has at the starting point of simulation. At any given time, the variable "energy" indicates the amount of energy in a node. The original energy value is passed on as an input argument. For each packet transmission, a node loses a particular amount of energy. As a result, the value of a node's initial energy is reduced and is known as the "Residual Energy". This residual energy is evaluated at different times by accessing the built-in variable "energy" in the find Energy method.

By considering the above parameters, we simulated 40 nodes with a total simulation time of 400 sec with the EECRM and our proposed NMABHM. The results obtained for the metrics are as follows:

a) *Average End-to-End delay:* Fig. 5 depicts the average end-to-end delay in milliseconds (m/s) experienced while transmitting data from mobile nodes to the Main Peer within 400 units of time. For simulation times up to 200 seconds, the EECRM and the NMABHM both produce outputs that are delayed by a similar amount. The delay in EECRM begins to increase after 200 units of time, but our NMABHM achieves respectable results in terms of delay.

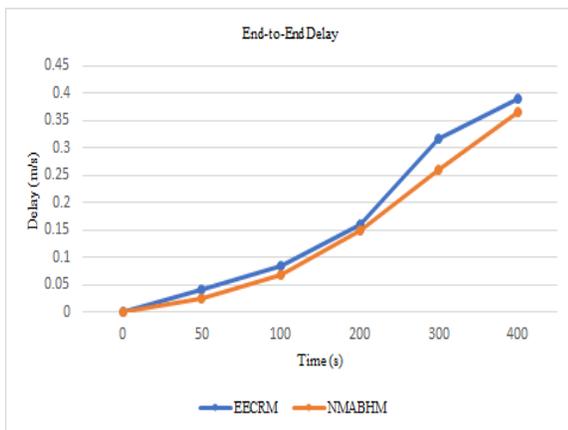


Fig. 5. Average End-to-End Delay in 400 Units of Time.

b) *Throughput:* It is defined as the number of successfully received packets in a unit time and is shown in Fig. 6. While the results of both approaches are identical at the beginning of the simulation, our proposed NMABHM strategy achieves a maximum throughput of 180 kbps (approximately) towards the end of the simulation.

c) *Average Energy consumption in NMABHM:* Fig. 7 depicts the average amount of energy spent by the MANET nodes during the communication process during a period of 400 units of time. This energy consumption is measured based

on the Optimal Route Selection using the TOPSIS Algorithm and the shortest Euclidean distance from the Super Peer to the Main Peer. During the simulation, it can be observed that the energy requirements of the existing EECRM approach is high than our suggested NMABHM approach. The energy consumption is reasonable in the midst of the execution, but at the end of the process, NMABHM has demonstrated reasonable energy usage.

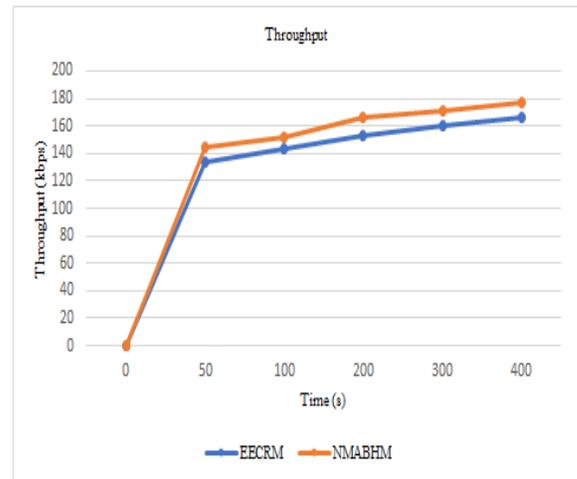


Fig. 6. Throughput in 400 Units of Time.

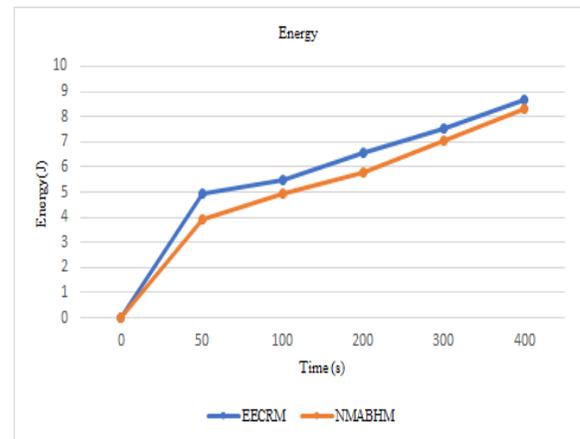


Fig. 7. Average Energy Consumption in 400 Units of Time.

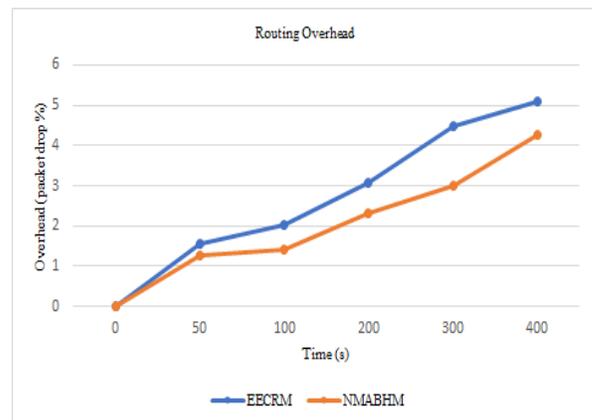


Fig. 8. Routing Overhead (RO) in 400 Units of Time.

d) *Routing Overhead*: The total number of routing packets needed for network communication is "Routing Overhead". Fig. 8 represents the routing overhead caused due to the route failure and route establishment of nodes in MANET. It is clear from the graph that there is a significant variation in the routing overhead results. The packet dropping rate is identical at the beginning of the simulation, but as the simulation progresses, the EECRM causes considerably more packet losses when compared to the NMABHM, resulting in a higher routing overhead when compared to our proposed work.

The experimental results show that the End-to-End delay in our proposed work has shown approximately 25% improvement than the compared EECRM. Also, there is 25-30% of energy gains, the Overhead is improved by approximately 10-15% and the throughput is achieved almost by 80-90%. The experimental results indicate that our suggested work can significantly minimize energy consumption in the event of a network link failure.

VI. CONCLUSION AND FUTURE WORK

In this work, we proposed a Node Monitoring Agent-based Handover Mechanism (NMABHM) in CA-MANET that helps in reducing the high energy consumption which occurs due to connectivity loss between the Super Peer and Indirect Peer. A Cluster-based CA-MANET architecture is suggested for effective use of bandwidth and energy in the network. We employ a Node Monitoring Agent (NMA) that acts as a Supervising node in the network to indicate the handover issue which can be effectively solved by adapting a multi-attribute decision-making technique called TOPSIS for Optimal route selection in MANET. The Simulation is done using an NS-2 simulator and the experimental results prove that the NMABHM mechanism helps in handling the Handover situation and lowering the energy consumption in the CA-MANET efficiently when compared to the existing EECRM mechanism. Simulations of our proposed work have been done with fewer nodes and with the assumption that cluster heads are static. By considering the cluster head's dynamic nature and scaling the simulation to include a larger number of nodes, we may expect our suggested technique to find many more practical applications in future work.

REFERENCES

- [1] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Commun. Surv. Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.
- [2] B. Manale and T. Mazri, "5G, vehicle to everything communication: Opportunities, constraints and future directions," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 6, pp. 1089–1095, 2020.
- [3] V. K. Quy, N. T. Ban, V. H. Nam, D. M. Tuan, and N. D. Han, "Survey of recent routing metrics and protocols for mobile Ad-hoc networks," *J. Commun.*, vol. 14, no. 2, pp. 110–120, 2019.
- [4] M. Ichaba, F. Musau, and S. N. Mwendia, "Performance effects of algorithmic elements in selected MANETs routing protocols," *Adv. Sci. Technol. Eng. Syst.*, vol. 5, no. 3, pp. 62–71, 2020.
- [5] Q. B. Hani and J. P. Dichter, "Energy-efficient service-oriented architecture for mobile cloud handover," *J. Cloud Comput.*, vol. 6, no. 1, 2017.
- [6] S. Mumtaz, K. Mohammed, S. Huq, and J. Rodriguez, "Direct mobile-to-mobile communication: Paradigm for 5G," *IEEE Wirel. Commun.*, vol. 21, no. 5, pp. 14–23, 2014.

- [7] T. Alam, "Middleware Implementation in Cloud-MANET Mobility Model for Internet of Smart Devices," *Int. J. Comput. Sci. Netw. Secur.*, vol. 17, no. 5, pp. 86–94, 2017.
- [8] J. Li, X. Li, Y. Gao, Y. Gao and R. Zhang, "Dynamic Cloudlet-Assisted Energy-Saving Routing Mechanism for Mobile Ad Hoc Networks," in *IEEE Access*, vol. 5, pp. 20908-20920, 2017.
- [9] N. D. Han, Y. Chung, and M. Jo, "Green data centers for cloud-assisted mobile ad-hoc network in 5G," *IEEE Network*, vol. 29, no. 2, pp. 70-76, 2015.
- [10] V. K. Quy, L. N. Hung, and N. D. Han, "CEPRM: A cloud14-assisted energy-saving and performance-improving routing mechanism for MANETs," *J. Commun.*, vol. 14, no. 12, pp. 1211–1217, 2019.
- [11] H. Yao, C. Bai, C. Hu, D. Zeng, and Q. Liang, "Survey on Mobile Data Offloading [J]," *Comput. Sci.*, vol. 41, no. 11A, pp. 182–186, 2014.
- [12] Umaphathyusha, B.V.S. & Babu, K., "A feasible rebroadcast system for lessening routing overhead in manets". *International Journal of Pharmacy and Technology*, vol. 8, no. 12, pp. 22314-22321, 2016
- [13] Maliki A, Owens G, Bruce D, "Combinig AHP and TOPSIS Approaches to Support Site Selection", *IP-CBEE*, vol. 37, pp. 1-8, 2012.
- [14] Moghadas, M., Asadzadeh, A., Vafeidis, A.T., Fekete, A., & Kötter, T, "A multi-criteria approach for assessing urban flood resilience in Tehran, Iran", *International Journal of Disaster Risk Reduction*, 2019.
- [15] Q. Pham et al., "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," in *IEEE Access*, vol. 8, pp. 116974-117017, 2020.
- [16] V. R. Verma, D. P. Sharma and C. S. Lamba, "Improvement in QoS of MANET Routing by finding optimal route using Mobile Agent paradigm and Intelligent Routing Decision using Fuzzy Logic Approach," *International Conference on Computing, Power and Communication Technologies (GUCON)*, pp. 725-730, 2019.
- [17] V. K. Quy, N. D. Han, and N. T. Ban, "A_WCETT: A high-performance routing protocol based on mobile agent for mobile ad hoc networks in 5G," *Journal on Infor. Techn. & Comm.*, vol. 17, no. 37, pp. 14-21, 2017. (in Vietnamese).
- [18] V. K. Quy, N. D. Han, and N. T. Ban, "An Advanced Energy Efficient and High Performance Routing Protocol for MANET in 5G". *J. Commun.*, vol. 13, no.12, pp. 743-749, 2018.
- [19] Zhang, F., & Yang, G, "A Stable Backup Routing Protocol for Wireless Ad Hoc Networks" in *Sensors (Basel, Switzerland)*, vol. 20, no. 23, pp. 6743, 2020.
- [20] Fatemeh Sarkohaki, Reza Fotohi and Vahab Ashrafian, "An Efficient Routing Protocol in Mobile Ad-hoc Networks by using Artificial Immune System" *International Journal of Advanced Computer Science and Applications(IJACSA)*, vol. 8, no. 4, 2017.
- [21] L. Li, C. Li, and P. Yuan, "An Energy Level Based Routing Protocol in Ad Hoc Networks," *Wirel. Pers. Commun.*, vol. 81, no. 3, pp. 981–996, 2015.
- [22] M. Tawalbeh, A. Eardley, and L. Tawalbeh, "Studying the Energy Consumption in Mobile Devices," *Procedia Comput. Sci.*, vol. 94, no. *MobiSPC*, pp. 183–189, 2016.
- [23] S. S. Muratchaev, A. S. Volkov, V. S. Martynov, and I. A. Zhuravlev, "Application of Clustering Methods in MANET," *Proc. 2020 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. EIConRus 2020*, no. 19, pp. 1711–1714, 2020.
- [24] G.Y. Park, H. Kim, H.W. Jeong, H.Y. Youn, "A novel cluster head selection method based on k-means algorithm for energy efficient wireless sensor network", *Proc. - 27th Int. Conf. Adv. Inf. Netw. Appl. Work. WAINA*, pp. 910–915, 2013.
- [25] Xiaonan, W. and Shan, Z., " Research on mobility handover for IPv6-based MANET". *Trans. Emerging Tel Tech*, vol. 25, no.7, pp. 679-691, 2014.
- [26] D. Li, X. Li, and J. Wan, "A cloud-assisted handover optimization strategy for mobile nodes in industrial wireless networks," *Comput. Networks*, vol. 128, no.6, pp. 133–141, 2017.
- [27] H. Abdulwahid, B. Dai, B. Huang and Z. Chen, "Energy-Efficient and Reliable Routing for Mobility Prediction-Based MANETs," *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, no. 12, pp. 43-51, 2015.

- [28] H. Riasudheen, K. Selvamani, S. Mukherjee, and I. R. Divyasree, "An efficient energy-aware routing scheme for cloud-assisted MANETs in 5G," *Ad Hoc Networks*, vol. 97, no. 10, pp. 102021, 2020.
- [29] Gharsallah, Amina, Zarai, "MIH/SDN-Based Vertical Handover Approach For 5G Mobile Networks", *Journal of Information Science and Engineering*, vol. 35, no.5, pp. 1161–1172, 2019.
- [30] Lai, Wei K., Chin-Shiuh Shieh, Fu-Sheng Chou, Chia-Yu Hsu, and Meng-Han Shen, "Handover Management for D2D Communication in 5G Networks" *Applied Sciences*, vol. 10, no. 12, pp. 4409, 2020.
- [31] Priyambodo, Tri K., Danur Wijayanto, and Made S. Gitakarma, "Performance Optimization of MANET Networks through Routing Protocol Analysis" *Computers*, vol. 10, no. 1: 2, 2021.

Design of an Intelligent Hydroponics System to Identify Macronutrient Deficiencies in Chili

Deffa Rahadiyan¹, Sri Hartati^{2*}, Wahyono³
Department of Computer Science and Electronics
Universitas Gadjah Mada
Special Region of Yogyakarta, Indonesia

Andri Prima Nugroho⁴
Departement of Agricultural and Biosystems Engineering
Universitas Gadjah Mada
Special Region of Yogyakarta, Indonesia

Abstract—Nutrient contents are important for plants. Lack of macronutrients causes plant damage. Several macronutrient deficiencies exhibit similar visual characteristics that are difficult for ordinary farmers to identify. Collaboration between Computer Vision technology and IoT has become a non-destructive method for nutrient monitoring and control, included in the hydroponic system. Computer vision plays a role in processing plant image data based on specific characteristics. However, the analysis of one characteristic cannot represent plant health. In addition, knowing the percentage of macronutrient deficiencies is also needed to support precision agriculture systems. Therefore, we propose a Multi Layer Perceptron architecture that can perform multi-tasks, namely, identification and estimation. In addition, the optimal architecture will also be sought based on the characteristics of the combination of three features in the form of texture, color, and leaf shape. Based on analysis and design, our proposed model has a high potential for identifying and estimating macronutrient deficiency at the same time as well and can be applied to support precision agriculture in Indonesia.

Keywords—Multi Layer perceptron; internet of things; feature combination; leaf image; nutrient deficiency

I. INTRODUCTION

The chili plant is a high economic value of the horticultural plant in Indonesia. However, the production level is lower than the consumption level, with inflation of 0.20% to 0.55% in 2019 [1]. The thing that causes low production is limited land due to the transition of the farming areas to settlements. Another thing is crop failure due to erratic weather changes, pest attacks, and plant diseases[2]. A System that can be applied to limited land in uncertain weather is hydroponics [3].

Hydroponics is a farming system that emphasizes the fulfillment of plant nutrients [4]. A plant needs macro and micronutrients to grow and develop [5]. Macronutrients include N, P, K, Ca, Mg, S (> 1000 mg/kg dry matter) and micronutrients include Iron, Mn, Zn, Cu, Cl, B, and Mo (<100 mg/kg dry matter)[6]. Inappropriate nutrient content causes plants to have macronutrient deficiency. Nutrient deficiency is more easily observed in the leaves [7], [8]. The symptoms in leaves include marginal, interveinal, and uniform chlorosis, distorted edges, reduction in the size of the leaf, necrosis, etc. [9]. However, identifying macronutrient deficiency is difficult for ordinary farmers because several nutrients show similar characteristics [10].

Technology in agriculture is an important part of supporting the application of industry 4.0 in Indonesia [11]. One of its applications is a monitoring and control system in intelligent hydroponics using the Internet of Things (IoT) [12]. One thing that needs to be monitored is the plant's health condition. Computer Vision is one of the technologies that can help the farmer to check the condition of plants [13], [14]. Collaboration between IoT technology and computer vision produces an automatic hydroponic system that can find out the condition of the plant and the solution.

Computer vision technology utilizes image data for analysis. Several images color model have been used, one of them is an RGB image that works like the human eye which is sensitive to the red, green, and blue light bands [15], [16]. Identification of macronutrient deficiencies using RGB leaves images has been carried out. Coffee, tomato, chili, cucumber, etc. have been analyzed using texture and color using K-Nearest Neighbors (KNN), Artificial Neural Network (ANN), Naïve Bayes [17], Multi Layer Perceptron (MLP), and Convolutional Neural Network (CNN) with various architecture [18], [19]. MLP gives a promising result. However, most of these studies only determined the type of nutrient deficiency [20]. To support a precise agricultural system, the percentage of deficiency is important so nutrient solutions accordance with plant needs [21], [22].

The percentage of macronutrient deficiency estimation using plant RGB images has been carried out [16], [23]. Color characteristics in plants have been used as in wheat using Multivariate Linear Regression, Genetic Algorithm, Back Propagation-ANN, and KNN [23], [24]. Texture features have also been developed using the Support Vector Machine (SVM) algorithm[25]. Several studies have also utilized deep learning such as in chili plants using Recurrent Convolution Neural Network (RCNN) [1]. Most of them only use one feature for estimation, so the resulting model is not robust for all types of macronutrients [26]. It because of each macronutrient shows different visual characteristics.

The proposed study is image-processing-based to identify and estimate the percentage of macronutrient deficiencies in a hydroponics environment. There are several phases to determine the lack of macronutrients such as image acquisition, preprocessing, segmentation, feature extraction, identification, and estimation task. The smart hydroponic system is assisted by IoT so that the process of image data acquisition is carried out automatically. Image data is crop

*Corresponding Author.

data in real conditions so, preprocessing such as histogram equalization is required. Then, the chili plant objects are separated from a complex background using a segmentation technique. One of the challenges faced at the segmentation stage is separating overlapping leaves images. A combination of three features in the form of color, texture, and leaf shape is proposed to support precision agriculture. Careful selection of features must be conducted to obtain accurate models. The main contribution in this paper is identification and estimation in a multi-layer perceptron architecture based on combination features. The study aims to identify and estimates five types of plant conditions such as healthy, potassium deficiency, calcium deficiency, magnesium deficiency, and Sulphur deficiency.

II. RELATED WORK

The study of macronutrient identification and estimation has been done from destructive to non-destructive methods [5], [16]. The destructive method is tested in the laboratory but it is costly and time-consuming[27]. Then, a monotonous and long duration of work would raise a human error [28]. Therefore, non-destructive is needed.

Internet of Thing is one of the non-destructive methods that have been used [29]. IoT can be used to monitor and control physical phenomena around agricultural environments such as temperature, light intensity, pH, and others [30], [31], [33]. Not only observing, but IoT also helps to determines the health condition of plants based on the image data [38]. However, the visual condition of the plant is not considered, even though the condition of the plant is important. Table I shows visual characteristics of several macronutrient deficiencies. A sensor that can perform visual plant condition is camera sensors that usually used for monitoring the agricultural environment [39].

The camera is useful for capturing plant images in a hydroponic environment [34]. Some of the plants that have been observed are tomato, chili, tobacco, spinach, okra, and others [28], [35], [36]. Digital image processing is used to analyze the captured images to obtain plant conditions. Plant conditions that will be analyzed are the type of macronutrient deficiency and the percentage. These stages are preprocessing and augmentation, segmentation, feature extraction, to identification and estimation tasks [22], [37].

Preprocessing uses to improve image quality[48]. Several preprocessing types are an enhancement, resizing, histogram equalization, and others[7]. Histogram equalization has been used to cope with different exposures on captured images [39], [40]. Then, data augmentation is a process for data enrichment. One example of data augmentation is processing rotation and blurring [41]. The study [42], [43] has proven that image augmentation so that each class has the same amount of data increases the accuracy of the model.

Segmentation is process to separate objects from the background, even a complex background [9], [16]. Several methods have been applied for segmentation, from K-Means Clustering, Genetic Algorithm with DSELM, Otsu, masking

green, Fuzzy C-Means (FCM) method to thresholding [21], [22], [44]. However, thresholding and green masking methods cannot overcome overlapping leaves. FCM is a clustering method that can group data. In [45], it is proven that FCM can overcome overlapping objects.

Feature extraction can be done using several methods based on the information characteristics. Statistical features in the RGB, HSV, and YUV color models represent color information from objects [24], [32]. While some methods, such as GLCM, Sobel, etc., can represent texture and shape information on leaf objects [17]. Feature combination is important because each deficiency shows a different visual features[9]. There is a nutrient that shows color characteristics, and some do not. Therefore, a combination of 3 features such as color, shape, and leaf texture is needed. A combination GLCM, hue, and color histogram has been used to analyze maize plants [46]. Then, a combination of RGB and Sobel edge improves the accuracy [21], [35].

Several methods have been used to identify and estimate macronutrient deficiencies in plants, such as Rule-based method to deep learning in the different models [47], [48]. The rule-based, histogram, and RGB-based method were used to identify leaf deficiency [48], [49]. However, those methods cannot handle data with high variance. Learning methods such as supervised learning have also been used, such as MLP, ANN, KNN, SVM, and others [24], [50], [51]. MLP shows promising results for data in different lighting conditions [52]. It can be concluded that MLP is a method that can be used to process hydroponic data in a natural environment. Each method is only robust on certain macronutrients. It is proven in [26] study using the derivative function method. The first derivative obtains more accurate results for predicting N, P, and S. Nutrient such as Mg, and K in plants uses logarithmic transformation while Ir uses smoothed reflectance. Br and Mn use the first derivative, while Zn uses the second derivative. In addition, CNN with various architectures has been used. However, it is a black box method [2], [8], [53] so it can be challenging to analyze the effect of feature combinations on the resulting model. Table II shows a summary of some of the studies that have been mentioned.

TABLE I. KEY SYMPTOMS OF NUTRIENT DEFICIENCIES [19]

Type of Deficiencies	Characteristics of leaves plant			
	Colour	Shape	Texture	Part of Plant
(-Ca)	Healthy Green	Misshapen	curling leaf tip	Top
(-N)	Turning yellow	Ellipse	Smooth	Bottom
(-Mg)	Necrosis (cell injury), interveinal chlorosis	Ellipse/ Misshapen	Smooth	Bottom
(-K)	Brown in the edge	Misshapen	Curling	Top
Health	Green	Ellipse	Smooth	overall

TABLE II. RESEARCH ON IDENTIFICATION MACRONUTRIENT DEFICIENCY

Plant	Dataset	Preprocessing	Feature Extraction	Classifier	Result	Reference
Paddy	500 images for 5 categories	resize to 256x256 and Median Filter	RGB feature (0-255)	Rule based	improving the database and refining the rules can improve the accuracy.	[54]
Paddy	400 images for 5 categories	RGB to HSV	HSV color model	Rule based	86% of accuracy	[48]
Coffee	269 images for 4 categories	Segmentation using Otsu method	BSM and GLCM to extract Shape and Texture	KNN, Naïve bayes, NN	the best results associated with the GLCM for Iron	[17]
Wheat	360 images	Segmentation using DSELM and GA	Color	a number of DSELMs and committee with GA	single color features and their combinations are not suitable to estimate nitrogen	[22]
Okra Plant	231 images for 4 categories	Resizing image to 299 x 299 and data augmentation using ImageDataGenerator by Keras	Automated FE	Inception ResNet-v2 CNN	the amount of data for several deficiency only produces 2 classes	[35]
Banana	540 images for 4 categories	Data augmentation, RGB convert to YUV, HE, CIELAB, YCbCr and HSV.	Automated FE	Neural Network. VGG16 model.	The model has very high accuracy and minimal error by data augmentation	[42]
Tobacco	204 tobacco leaves for tested	RGB to HSV color space, image thresholding, morphological operation, and crop leaf image	Color analysis	Voting from each patch based on a set of thresholds	The proposed method was able to detect the defect and classify tobacco leaf with 91.667% accuracy	[28]
Eucalyptus	100 leaves	First and second Derivative, Leaf reflectance logarithmic transformation	Color feature	PLSR	each nutrient has high performance in different methods	[26]
Tomato	596 images	596 images of 3024 x 4032 px size were stored in the dataset	a set of convolutional layers as the feature extraction part,	CNN+AHN	CNN+AHN is able to estimate with an accuracy of 95.57% outperforming the literature	[8]
Olive tree	4000 images for 2 categories	RGB to HSV color space, masking	YGB, percentage of RGB value	ML, decision tree and Naïve Bayes models	machine learning model showed the best results with 97% accuracy.	[55]
Multi Plant	484 for training and 13 for testing	There is no pre-processing	Automated FE	CNN	the model only search for colors in the leaf images	[56]
Multi Plant	5000 images of different plants	preprocessing task using SMOTE	CNN is used to extract the patterns of leaf images	the use of an EM classifier to supplement the ELM-based classifier.	proposed modified CNN model obtains improved training and testing classification accuracy than other methods.	[29]
Rice	8911 images for 5 categories	Segmentation using Mean Shift image. Then, RGB image is convert to grayscale for other color model such as YCbCr and HSI	Color using Ybr and HSI value, shape using traditional shape feature such as area, roundness, shape complexity	CNN, Deep learning CNN + SVM, AlexNet, VGG	CNN with image segmentation shows better result than without segmentation. But, CNN +SVM best result.	[57]
Vigna Mungo	4088 images for 7 categories	Flipping and resizing of the image,	taking the advantages of the deep pre-trained model using ResNet50	CNN, SVM, and MLP	MLP achieved superior performance than SVM and logistic regression by the accuracy of 88.33 %.	[14]

Based on some of the literature reviews above, it can be concluded that IoT and Computer Vision-based in smart hydroponic farming can help farmers find outcrop conditions

quickly. In addition, methods for analyzing image data are essential, such as MLP, which can process image data with different lighting. Then, MLP performs the best result with 88,33 accuracy than logistic regression and SVM in [14].

Several features can be used to analyze, such as histogram value, HSV statistic value, texture using GLCM, and others. But, several researches above shows that the types of features analyzed from the image data need to be combined to improve the quality of the model so that the model can be robust across various types of macronutrients.

III. PROPOSED SYSTEM

This section discusses intelligent hydroponics system workflow and Computer Vision in Hydroponics systems. The discussion of the two parts of this model answers the research question.

A. Smart Hydroponics System Workflow

Overall, there are three steps in a hydroponics system such as data acquisition, data analysis, and user area. Fig. 1 presents a schematic of the proposed model. Based on Fig. 1 below, Part A is the data acquisition stage. Part B contains a process to identify and estimate macronutrient deficiency in the plant. Then, part C is an interface that connects farmers with hydroponic systems so plants and agricultural environments can be observed and controlled well. The hydroponic system runs automatically. The detail of each part is discussed below:

- **Data Acquisition:** There are two types of data in the hydroponic system: agricultural environmental and plant conditions data. The environment data includes light intensity, humidity, temperature, and water levels. While plant conditions are plant image data. Each sensor acquires different data. Then via Raspberry Pi, data is sent to the server using Wi-Fi and other protocol.

Image plant data is taken in the morning and afternoon to avoid distraction by lighting conditions. Image data is taken using a camera sensor with a certain distance.

- **Data Analysis:** Data is grouped into a database according to their needs. Plant image is used to test the identification and estimation model that has been built using digital image processing and machine learning. The result is plant condition and their percentages. These results are used as input on the Nutrient Control to give nutrient solutions according to plant needs.
- **Web Development:** Hydroponic runs automatically, but each process can be observed on the web development interface that can be accessed through computers or smartphones. So, users can observe the ongoing process and control the agricultural environment.

B. Identification and Estimation Process

Image data are analyzed using digital image processing. Fig. 2 shows a flowchart of identification and estimation of macronutrient deficiency. The stages of digital image processing are divided into five stages. The first stage is preprocessing using histogram equalization and extraction of three types of features. Features are selected and combined with high considerations so that the analyzed features can represent the condition of the plant, then, training to determine the percentage of visual characteristics match. The percentage of visual characteristics becomes the input for the MLP training stage. The output is an appropriate identification and estimation model.

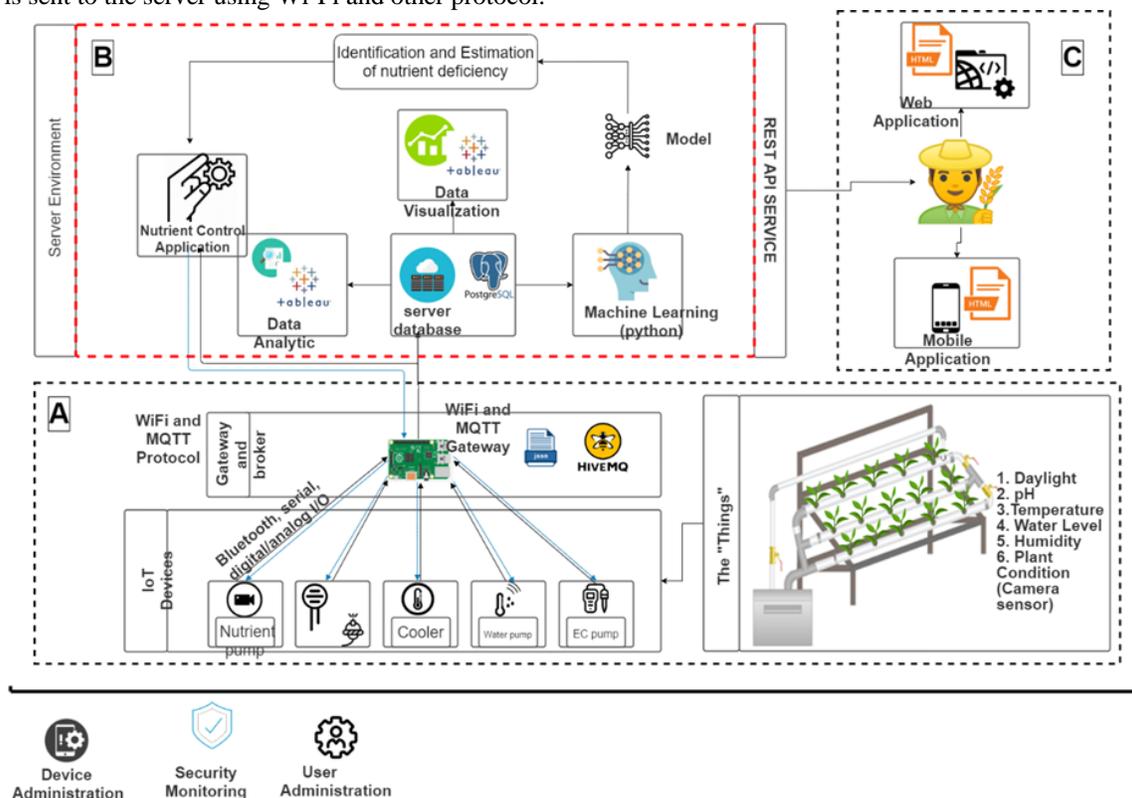


Fig. 1. Smart Hydroponics Design System.

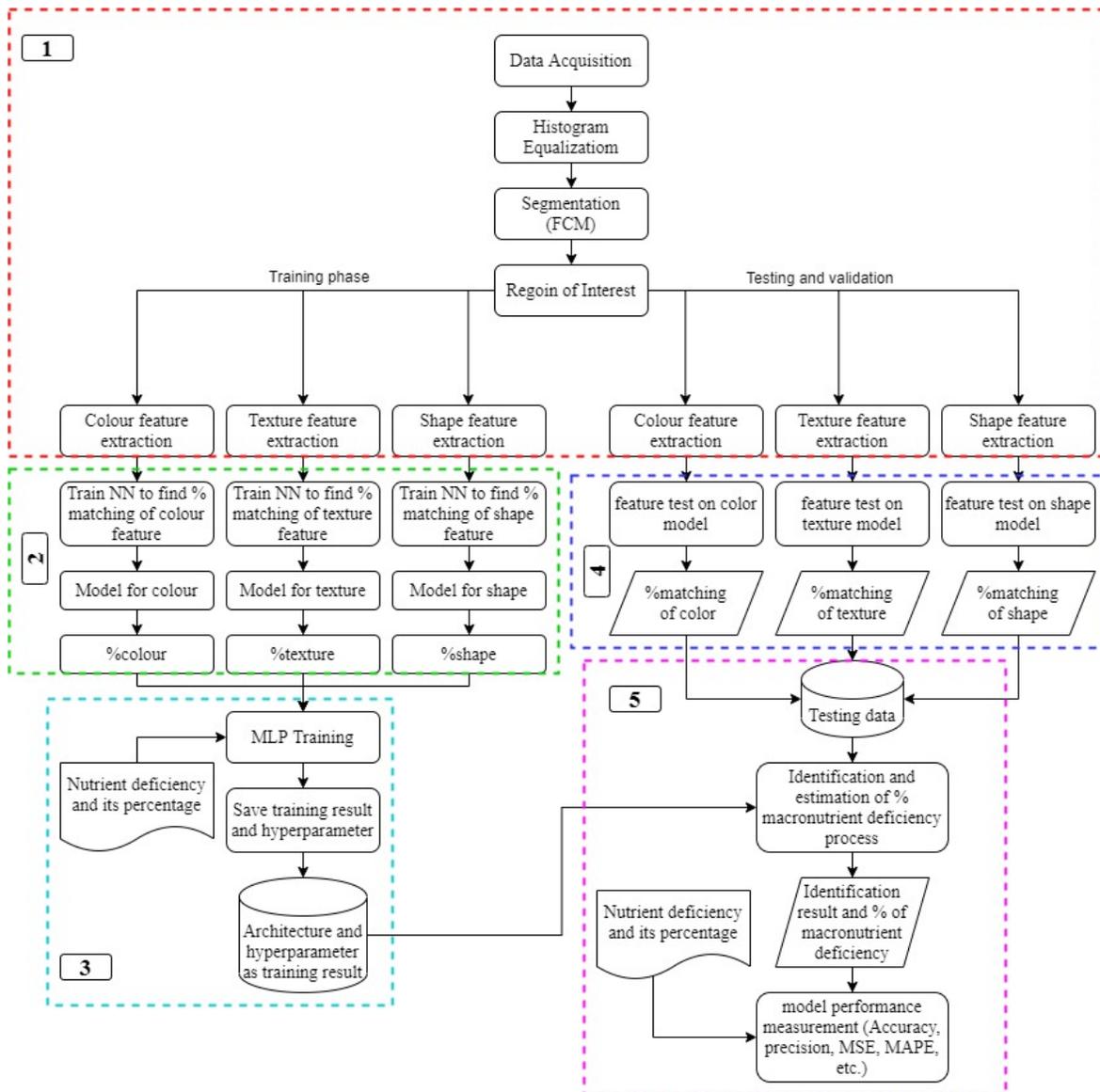


Fig. 2. Identification and Estimation Macronutrient Deficiency Steps.

In this proposed method, there are two types of training data. The first train is to determine the percentage of matches based on visual characteristics. Then, the second train uses the percentage of matching data obtained from the first training for identification and estimation tasks. Our model develops a system that can identify macronutrient deficiencies and estimate the percentage of deficiency at one time. In addition, we are also looking for the suitable MLP architecture based on the combination of three features used, namely color, shape, and leaf texture. The model is evaluated using two different types of evaluators based on their tasks. The last is the resulting model will test in a natural hydroponic system environment that has been build. Each stage is described in detail in the next section.

IV. EXPERIMENTAL RESULT AND DISCUSSION

This section discusses the detail process of image processing in hydroponic systems. The camera acquires the

image data in natural environment. Fig. 3 shows some results of image acquisition of chili plants. Healthy plants are given different treatments by reducing certain macronutrients. After 1 week, the plants will show visual characteristics. The planned amount of data taken is 500 for each class. The data includes data with a percentage label of the deficiency.



Fig. 3. Image Plant Data in Real Condition.

1) *Preprocessing and feature extraction*: Each training and testing data are treated the same at this stage. The result of the acquisition is an RGB image with different lighting. Histogram equalization is applied to equalize the different exposures in the image. Where the results of the conversion of RGB to HSV images are transformed into histograms so that they can be normalized using the Cumulative Distribution Function (CDF) as shown in (1) where n_{r_j} is the histogram value and n is the total [51]. Then the intensity transformation is carried out from the input image r_k to s_k as shown in (2) with a pixel of level L .

$$cdf(j) = \sum_{j=0}^k \frac{n_{r_j}}{n} \quad (1)$$

$$s_k = T(r_k) = \text{round}(cdf(r_k) \times (L - 1)) \quad (2)$$

The result of Histogram equalization shown on Fig. 4 It will be segmented to separate objects from complex backgrounds using FCM. FCM is a clustering method that can be used for abstract data. In this case, FCM is used to overcome overlapping leaves. The basic idea of FCM is to divide n pieces of data into non-unique sets to improve cluster data based on membership degrees, where the membership degrees are real numbers in the range [0,1] [52]. To select the leaves that are not overlapping ROI is applied. Segmentation and ROI result are several single leaf images whose leaves do not overlap as shown on Fig. 5.

Feature extraction is performed for each single leaf image. In this study, three features used are: shape, texture, and leaf color. These traits were chosen based on the leaves visual characteristics. Each feature is processed using a different method. Color feature extraction using HSV values, texture feature extraction using Gray Level Co-occurrence Matrix (GLCM), and shape feature extraction using Canny edge detection with Freeman Chain code. The output of color feature extraction is Mean value in (3), Standard deviation in (4), and skewness in (5) for each H, S, and V [48]. Where μ is Mean, σ is Standard Deviation, M is image dimension based i -th pixel, N is the total number of j -th, and I_{ij} is value of the j -th pixel of the image at the i -th color channel.

$$\mu = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N I_{ij} \quad (3)$$

$$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (I_{ij} - \text{Mean})^2} \quad (4)$$

$$\text{Skewness} = \sqrt[3]{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N (I_{ij} - \text{Mean})^3} \quad (5)$$

While the GLCM output is two parameters, such as Angular second moment (ASM) in (6) and Inverse Difference Moment (IDM) in (7) [17]. In shape feature extraction, it is still a chain code. Chain code can be converted into 15 normalized Elliptic Fourier Descriptors based on the Chain code result.

$$\text{Kontras} = \sum_{n=1}^L n^2 \{ \sum_{|i-j|=n} \text{GLCM}(i, j) \} \quad (6)$$

$$\text{IDM} = \sum_{i=1}^L \sum_{j=1}^L \frac{(\text{GLCM}(i, j))^2}{1+(i-j)^2} \quad (7)$$

2) *Training to get matching percentage of visual characteristics*: The results of each feature extraction are trained using NN-Backpropagation to check the matching percentage of characteristics. The output of NN-Backprop is the percentage matching characteristics which grouped into specific types, as shown in Table III. For example, for Nitrogen, The possible matching percentage is shown in Table III. The output of NN-Backprop is input into the MLP architecture, so the type of nutrient deficiency and its percentage are known.

3) *Training to get identification and estimation model using mlp*: The model proposed in this study is an MLP architecture that can perform estimation and identification simultaneously. The model has two output types. They are numerical predictions and class predictions. Fig. 6 shows the proposed MLP architecture.

The type of MLP developed is fully connected. Each neuron is accumulated using a certain activation function where W_i is the weight of the i -th data, X_i is the i -th data, the value of b is the bias, and y is the output. Neuron input consists of nine values of chili plant matching percentage normalized with a range of [0-1]. After normalization, initialized weights and biases are given for each input neuron to the hidden layer with values $w_{11}, w_{12}, \dots, w_{xx}$, and $b_{11}, b_{12}, \dots, b_{xx}$ [19]. The decision layer has three neurons, where two neurons are for identification and one neuron is for estimation. The MLP output representation uses a binary output with the number of classes is 2^n where n is the number of neurons.

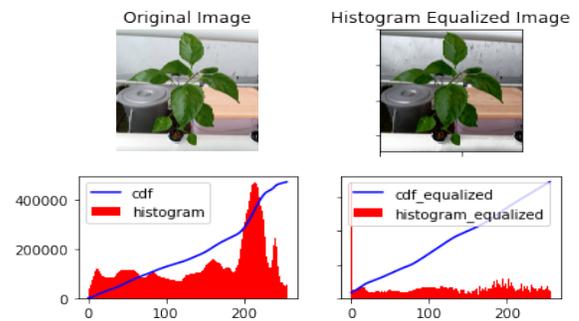


Fig. 4. Histogram Equalization Result.

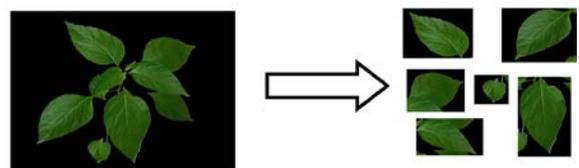


Fig. 5. Segmentation and ROI Result.

TABLE III. TYPE OF VISUAL CHARACTERISTICS

Colour				Texture			Shape	
Dark green	Yellow	Brown	Green	smooth	wavy	hole	Normal	Mishapen
10%	75%	5%	10%	80%	15%	5%	90%	10%

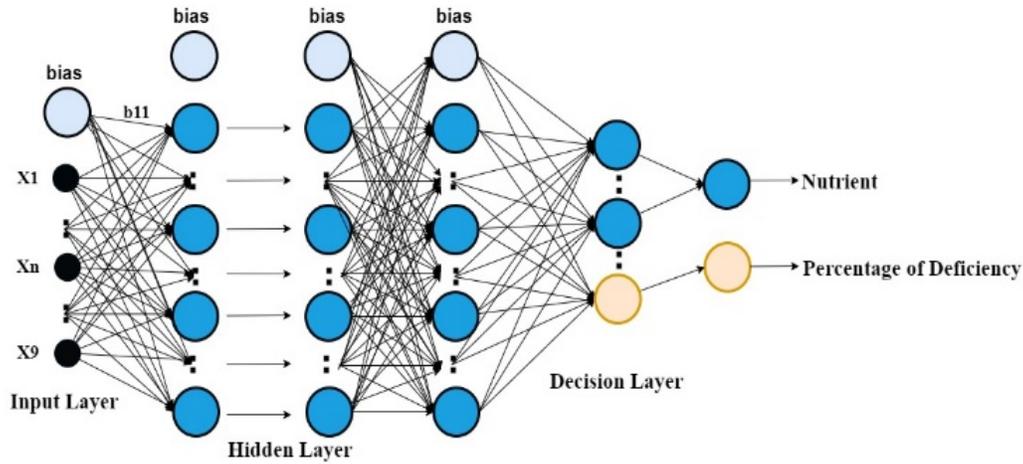


Fig. 6. Multi Layer Perceptron Architecture Design.

The activation function in the hidden layer is different from the output layer. Each neuron in the hidden layer uses the Rectified Linear Unit (ReLU). ReLU makes a limit on the number zero, if $x \leq 0$ then $x = 0$ and if $x > 0$ then $x = x$ as shown in (8)[14]. Meanwhile, in the decision layer, two different activation functions are used. A numerical prediction has single node and SoftMax activation function[5]. SoftMax not only maps the outputs to the range [0,1] but also maps each output with the total sum is 1. Therefore, SoftMax's output is suitable for the binary classification problem [0,1] as shown in (9). The Sigmoid activation function as shown in (10) is used for the estimation[58]. The sigmoid output normalized to obtain the percentage of macronutrient deficiency in the range 0%–100%.

$$f(x) = \max(0, x) \quad (8)$$

$$Softmax = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (9)$$

$$\phi = \frac{1}{1+e^{-y}}, \text{ where } y = (\sum_{i=1}^n W_i X_i + b) \quad (10)$$

In the machine learning method, the loss function is used to optimize the model during training so that the error is minimum. In identification task, the loss function that will be used is Categorical Cross Entropy that can be shown in (11) where L_{CE} is the loss category entropy, T_i is the target value and S_i is the result of the Softmax. A good model has L_{CE} close to 0. However, estimation task uses Mean Square Error (MSE) [59].

$$L_{CE} = -\sum_{i=1} T_i \log(S_i) \quad (11)$$

The number of hidden layers used affects the learning process. More hidden layers are used, the deeper features are studied. In the proposed model, the exact number of hidden layers and nodes of MLP is sought so that the optimal architecture is produced based on the three features combination. The effect of the number of epochs, learning rate, and others are considered.

4) *Find match percentage of testing data:* The test data that has been through the preprocessing process is searched for the matching percentage of the visual characteristics using

the NN-Backprop model from the previous stage. The expected output is to know each matching percentage of the visual characteristics.

5) *Identification and estimation macronutrient deficiency:* The matching percentage of its characteristics data is tested on the MLP model to know plant condition and its percentage. An evaluation is carried out to find out the model's performance against the test data. The identification task is evaluated using the confusion matrix to get the accuracy and precision based in (12) and (13) [9,14]. The rule of a confusion matrix is shown in Table IV.

TABLE IV. CONFUSION MATRIX REPRESENTATION [60]

		Prediction Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

While the estimation task is evaluated by calculating the error using Mean Square Error (MSE) and Mean Absolute Percentage Error (MAPE). Predictive result are good if the MAPE value is less than 10% [58]. Meanwhile, for MSE using a gradient-based method, a lower value makes a better prediction. The formula for calculating MAPE and MSE can be observed in (14) and (15) where Y_t is the actual value of period t , Y'_t is the forecast value of period t , and n is the number of periods.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (12)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (13)$$

$$MSE = \frac{1}{n} (\sum_{i=1}^n (Y_t - Y'_t)^2) \quad (14)$$

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|Y_t - Y'_t|}{Y_t} \quad (15)$$

The intelligent hydroponic system can identify nutrient deficiencies and provide solutions automatically. Based on the analysis and literature study that has been done, the proposed

model potential is to be applied. The MLP architecture performs two different tasks, namely, classification and regression. In addition to developing a multi-task architecture, the number of hidden layers and nodes will find based on the combination of the three features used.

V. CONCLUSION

This research is in the stage of data collection and software development. Intelligent hydroponics hardware has been built and is still being evaluated based on system requirements. This proposed model answers research questions about identifying and estimating macronutrient deficiency in intelligent hydroponics systems. First, identification and estimation are faster and more precise so that the provision of plant considers the needs of plants. Second, three features combination such as color, shape, and leaf texture is applied to increase information so that the system can identify the right type of nutrient. Third, build a model that can perform both identification and estimation tasks. Fourth, it analyzes more than one type of nutrients to be applied in the real environment. In the future, this model can be implemented in the natural environment of intelligent hydroponics systems in Indonesia. Limitations still exist because the model is only implemented in chili hydroponics plants.

ACKNOWLEDGMENT

This work is funded by the Ministry of Research and Technology of the Republic of Indonesia in the PMDSU program. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

- [1] A. R. Bahtiar, Pranowo, A. J. Santoso, and J. Juhariah, "Deep Learning Detected Nutrient Deficiency in Chili Plant," 2020 8th Int. Conf. Inf. Commun. Technol. ICoICT 2020, 2020, doi: 10.1109/ICoICT49345.2020.9166224.
- [2] L. C. Ngugi, M. Abelwahab, and M. Abo-Zahhad, "Recent advances in image processing techniques for automated leaf pest and disease recognition – A review," *Inf. Process. Agric.*, vol. 8, no. 1, pp. 27–51, 2021, doi: 10.1016/j.inpa.2020.04.004.
- [3] U. Nurhasan, A. Prasetyo, G. Lazuardi, E. Rohadi, and H. Pradibta, "Implementation IoT in System Monitoring Hydroponic Plant Water Circulation and Control," *Int. J. Eng. Technol.*, vol. 7, no. 4.44, p. 122, 2018, doi: 10.14419/ijet.v7i4.44.26965.
- [4] E. S. Putra, J. Jamaludin, and M. D. Djatmiko, "Comparison of Hydroponic System Design for Rural Communities in Indonesia," *J. Arts Humanit.*, vol. 7, no. 9, pp. 14–21, 2018, doi: 10.18533/journal.v7i9.1490.
- [5] L. N and K. K. Saju, "Classification of Macronutrient Deficiencies in Maize Plant Using Machine Learning," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 6, pp. 4197–4203, 2018, doi: 10.11591/ijece.v8i6.pp4197-4203.
- [6] T. T. Tran, J. W. Choi, T. T. H. Le, and J. W. Kim, "A comparative study of deep CNN in forecasting and classifying the macronutrient deficiencies on development of tomato plant," *Appl. Sci.*, vol. 9, no. 8, 2019, doi: 10.3390/app9081601.
- [7] L. Kamelia, T. K. B. A. Rahman, H. Saragih, and R. Haerani, "The comprehensive review on detection of macro nutrients deficiency in plants based on the image processing technique," *Proc. - 2020 6th Int. Conf. Wirel. Telemat. ICWT 2020*, pp. 7–10, 2020, doi: 10.1109/ICWT50448.2020.9243623.
- [8] H. Ponce, C. Cevallos, R. Espinosa, and S. Guti'erez, "Estimation of Low Nutrients in Tomato Crops Through the Analysis of Leaf Images Using Machine Learning," *J. Artif. Intell. Technol.*, vol. 1, no. 2, 2021, doi: 10.37965/jait.2021.0006.
- [9] S. Jeyalakshmi and R. Radha, "a Review on Diagnosis of Nutrient Deficiency Symptoms in Plant Leaf Image Using Digital Image Processing," *ICTACT J. Image Video Process.*, vol. 7, no. 4, pp. 1515–1524, 2017, doi: 10.21917/ijivp.2017.0216.
- [10] N. Minni and N. Rehna, "Detection of Nutrient Deficiencies in Plant Leaves using Image Processing," *Int. J. Comput. Algorithm*, vol. 5, no. 2, pp. 84–87, 2016, doi: 10.20894/ijcoa.101.005.002.004.
- [11] Y. Setiawan, H. Tanudjaja, and S. Octaviani, "Penggunaan Internet of Things (IoT) untuk Pemantauan dan Pengendalian Sistem Hidroponik," *TESLA J. Tek. Elektro*, vol. 20, no. 2, p. 175, 2019, doi: 10.24912/tesla.v20i2.2994.
- [12] K. Kularbphetong, U. Ampant, and N. Kongroj, "An Automated Hydroponics System Based on Mobile Application," *Int. J. Inf. Educ. Technol.*, vol. 9, no. 8, pp. 548–552, 2019, doi: 10.18178/ijiet.2019.9.8.1264.
- [13] V. R. P. Marcelo and J. G. Lagarteja, "Corzea: Portable maize (Zea Mays L.) nutrient deficiency identifier," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 1049–1053, 2020.
- [14] K. A. Myo Han and U. Watchareeruetai, "Black Gram Plant Nutrient Deficiency Classification in Combined Images Using Convolutional Neural Network," 2020 8th Int. Electr. Eng. Congr. iEECON 2020, 2020, doi: 10.1109/iEECON48109.2020.229562.
- [15] M. Merchant, V. Paradkar, M. Khanna, and S. Gokhale, "Mango Leaf Deficiency Detection Using Digital Image Processing and Machine Learning," 2018 3rd Int. Conf. Conver. Technol. I2CT 2018, pp. 1–3, 2018, doi: 10.1109/I2CT.2018.8529755.
- [16] S. Kolhar and J. Jagtap, "Plant trait estimation and classification studies in plant phenotyping using machine vision – A review," *Inf. Process. Agric.*, no. xxxx, 2021, doi: 10.1016/j.inpa.2021.02.006.
- [17] M. Vassallo-Barco, L. Vives-Garnique, V. Tuesta-Monteza, H. I. Mejía-Cabrera, and R. Y. Toledo, "Automatic detection of nutritional deficiencies in coffee tree leaves through shape and texture descriptors," *J. Digit. Inf. Manag.*, vol. 15, no. 1, pp. 7–18, 2017.
- [18] C. Wang, Y. Ye, Y. Tian, and Z. Yu, "Classification of nutrient deficiency in rice based on CNN model with Reinforcement Learning augmentation," *Proc. - 2021 Int. Symp. Artif. Intell. its Appl. Media, ISAIAM 2021*, pp. 107–111, 2021, doi: 10.1109/ISAIAM53259.2021.00029.
- [19] U. Watchareeruetai, P. Noinongyao, C. Wattanapaiboonsuk, P. Khantiviriya, and S. Duangsrisai, "Identification of Plant Nutrient Deficiencies Using Convolutional Neural Networks," *iEECON 2018 - 6th Int. Electr. Eng. Congr.*, pp. 2018–2021, 2018, doi: 10.1109/IEECON.2018.8712217.
- [20] V. Aleksandrov, "Identification of nutrient deficiency in bean plants by prompt chlorophyll fluorescence measurements and Artificial Neural Networks," *bioRxiv*, no. June, p. 664235, 2019, doi: 10.1101/664235.
- [21] Q. Wang, X. Mao, X. Jiang, D. Pei, and X. Shao, "Digital image processing technology under backpropagation neural network and KMeans Clustering algorithm on nitrogen utilization rate of Chinese cabbages," *PLoS One*, vol. 16, no. 3 March 2021, pp. 1–24, 2021, doi: 10.1371/journal.pone.0248923.
- [22] S. B. Sulisty, D. Wu, W. L. Woo, S. S. Dlay, and B. Gao, "Computational Deep Intelligence Vision Sensing for Nutrient Content Estimation in Agricultural Automation," *IEEE Trans. Autom. Sci. Eng.*, vol. 15, no. 3, pp. 1243–1257, 2018, doi: 10.1109/TASE.2017.2770170.
- [23] Y. Song, G. Teng, Y. Yuan, T. Liu, and Z. Sun, "Assessment of wheat chlorophyll content by the multiple linear regression of leaf image features," *Inf. Process. Agric.*, no. xxxx, 2020, doi: 10.1016/j.inpa.2020.05.002.
- [24] A. Qur'ania, P. Harsani, T. Triastinurmatiningsih, L. A. Wulandhari, and A. A. S. Gunawan, "Color Extraction and Edge Detection of Nutrient Deficiencies in Cucumber Leaves Using Artificial Neural Networks," *CommIT (Communication Inf. Technol. J.)*, vol. 14, no. 1, p. 23, 2020, doi: 10.21512/commit.v14i1.5952.
- [25] M. Mirzaei, S. Marofi, M. Abbasi, and R. Karimi, "Nondestructive estimation of leaf nitrogen and chlorophyll contents in grapes using field hyper spectral data and support vector machines approach," no. December, 2019, [Online]. Available: https://www.researchgate.net/publication/338006025_Nondestructive_es

- timation_of_leaf_nitrogen_and_chlorophyll_contents_in_grapes_using_field_hyper_spectral_data_and_support_vector_machines_approach%0D.
- [26] L. F. R. Oliveira and R. C. Santana, "Estimation of Leaf Nutrient Concentration from Hyperspectral Reflectance in Eucalyptus using Partial Least Squares Regression," *Sci. Agric.*, pp. 1–10, 2020, doi: <http://dx.doi.org/10.1590/1678-992X-2018-0409>.
- [27] M. R. Motsara and R. N. Roy, *Guide to laboratory establishment for plant nutrient analysis, FOOD AND AGRICULTURE ORGANIZATION OF THE UNITED NATIONS Rome, 2008. 2008.*
- [28] A. Harjoko, A. Prahara, T. W. Supardi, I. Candradewi, R. Pulungan, and S. Hartati, "Image processing approach for grading tobacco leaf based on color and quality," *Int. J. Smart Sens. Intell. Syst.*, vol. 12, no. 1, pp. 1–10, 2019, doi: [10.21307/ijssis-2019-010](https://doi.org/10.21307/ijssis-2019-010).
- [29] R. Sathyavani, K. Jaganmohan, and B. Kalaavathi, "Detection of plant leaf nutrients using convolutional neural network based internet of things data acquisition," *Int. J. Nonlinear Anal. Appl.*, vol. 12, no. 2, pp. 1175–1186, 2021, doi: [10.22075/ijnaa.2021.5194](https://doi.org/10.22075/ijnaa.2021.5194).
- [30] A. Ullah, S. Aktar, N. Sutar, R. Kabir, and A. Hossain, "Cost Effective Smart Hydroponic Monitoring and Controlling System Using IoT," *Intell. Control Autom.*, vol. 10, no. 04, pp. 142–154, 2019, doi: [10.4236/ica.2019.104010](https://doi.org/10.4236/ica.2019.104010).
- [31] A. Dudwadkar, T. Das, S. Suryawanshi, R. Dolas, and T. Kothawade, "Automated Hydroponics with Remote Monitoring and Control Using IoT," vol. 9, no. 06, pp. 928–932, 2020.
- [32] S. Mashumah, M. Rivai, and A. N. Infansyah, "Nutrient Film Technique based Hydroponic System Using Fuzzy Logic Control," *Proceeding - 2018 Int. Semin. Intell. Technol. Its Appl. ISITIA 2018*, no. May 2019, pp. 387–390, 2018, doi: [10.1109/ISITIA.2018.8711201](https://doi.org/10.1109/ISITIA.2018.8711201).
- [33] R. Perwiratama, Y. K. Setiadi, and Suyoto, "Smart hydroponic farming with IoT-based climate and nutrient manipulation system," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 129–132, 2019, doi: [10.1109/ICAIIIT.2019.8834533](https://doi.org/10.1109/ICAIIIT.2019.8834533).
- [34] A. Mateo-Aroca, G. García-Mateos, A. Ruiz-Canales, J. M. Molina-García-Pardo, and J. M. Molina-Martínez, "Remote image capture system to improve aerial supervision for precision irrigation in agriculture," *Water (Switzerland)*, vol. 11, no. 2, pp. 1–21, 2019, doi: [10.3390/w11020255](https://doi.org/10.3390/w11020255).
- [35] L. A. Wulandhari et al., "Plant nutrient deficiency detection using deep convolutional neural network," *ICIC Express Lett.*, vol. 13, no. 10, pp. 971–977, 2019, doi: [10.24507/iceicel.13.10.971](https://doi.org/10.24507/iceicel.13.10.971).
- [36] M. C. F. Lima, A. Krus, C. Valero, A. Barrientos, J. Del Cerro, and J. J. Roldán-Gómez, "Monitoring plant status and fertilization strategy through multispectral images," *Sensors (Switzerland)*, vol. 20, no. 2, 2020, doi: [10.3390/s20020435](https://doi.org/10.3390/s20020435).
- [37] H. Tian, T. Wang, Y. Liu, X. Qiao, and Y. Li, "Computer vision technology in agricultural automation—A review," *Inf. Process. Agric.*, vol. 7, no. 1, pp. 1–19, 2020, doi: [10.1016/j.inpa.2019.09.006](https://doi.org/10.1016/j.inpa.2019.09.006).
- [38] Z. Chen, C. Ying, C. Lin, S. Liu, and W. Li, "Multi-View Vehicle Type Recognition with Feedback-Enhancement Multi-Branch CNNs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2590–2599, 2019, doi: [10.1109/TCSVT.2017.2737460](https://doi.org/10.1109/TCSVT.2017.2737460).
- [39] M. Abdul, H. Radhi, A. Musa, and O. Al-Hsniue, "Enhancement of the Captured Images Under Different Lighting Conditions Using Histogram Equalization Method," *Int. J. Latest Res. Sci. Technol. ISSN*, vol. 3, no. 3, pp. 25–28, 2015, [Online]. Available: <http://www.mnkjournals.com/ijlrst.htm>.
- [40] N. Ahmad and A. Hadinegoro, "Metode Histogram Equalization untuk Perbaikan Citra Digital," *Semin. Nas. Teknol. Inf. Komun. Terap.*, vol. 2012, no. Semantik, pp. 439–445, 2012, [Online]. Available: <http://publikasi.dinus.ac.id/index.php/semantik/article/view/185>.
- [41] K. Iizuka and M. Morimoto, "A nutrient content estimation system of buffet menu using RGB-D sensor," 2018 *Jt. 7th Int. Conf. Informatics, Electron. Vis. 2nd Int. Conf. Imaging, Vis. Pattern Recognition, ICIEV-IVPR 2018*, pp. 165–168, 2019, doi: [10.1109/ICIEV.2018.8641061](https://doi.org/10.1109/ICIEV.2018.8641061).
- [42] R. Guerrero, B. Renteros, R. Castaneda, A. Villanueva, and I. Belupu, "Detection of nutrient deficiencies in banana plants using deep learning," pp. 1–7, 2021, doi: [10.1109/icaacca51523.2021.9465311](https://doi.org/10.1109/icaacca51523.2021.9465311).
- [43] Z. Xu et al., "Using deep convolutional neural networks for image-based diagnosis of nutrient deficiencies in rice," *Comput. Intell. Neurosci.*, vol. 2020, 2020, doi: [10.1155/2020/7307252](https://doi.org/10.1155/2020/7307252).
- [44] D. Zermas, H. J. Nelson, P. Stanitsas, V. Morellas, D. J. Mulla, and N. Papanikolopoulos, "A Methodology for the Detection of Nitrogen Deficiency in Corn Fields Using High-Resolution RGB Imagery," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 4, pp. 1879–1891, 2021, doi: [10.1109/TASE.2020.3022868](https://doi.org/10.1109/TASE.2020.3022868).
- [45] Q. U. Safitri, A. F. Huda, and A. S. Awaludin, "SEGMENTASI CITRA MENGGUNAKAN ALGORITMA FUZZY c-MEANS (FCM) DAN SPATIAL FUZZY c-MEANS (sFCM)," *Kubik J. Publ. Ilm. Mat.*, vol. 2, no. 1, pp. 22–34, 2017, doi: [10.15575/kubik.v2i1.1471](https://doi.org/10.15575/kubik.v2i1.1471).
- [46] N. Sabri, N. S. Kassim, S. Ibrahim, R. Roslan, N. N. A. Mangshor, and Z. Ibrahim, "Nutrient deficiency detection in maize (*Zea mays* L.) leaves using image processing," *IAES Int. J. Artif. Intell.*, vol. 9, no. 2, pp. 304–309, 2020, doi: [10.11591/ijai.v9.i2.pp304-309](https://doi.org/10.11591/ijai.v9.i2.pp304-309).
- [47] M. A. Islam, N. Rahman Shuvo, M. Shamsoddjaman, S. Hasan, S. Hossain, and T. Khatun, "An Automated Convolutional Neural Network Based Approach for Paddy Leaf Disease Detection," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 1, pp. 280–288, 2021, doi: [10.14569/IJACSA.2021.0120134](https://doi.org/10.14569/IJACSA.2021.0120134).
- [48] M. V. Latte, S. Shidnal, and B. S. Anami, "Rule Based Approach to Determine Nutrient Deficiency in Paddy Leaf Images," *Int. J. Agric. Technol.*, vol. 13, no. 2, pp. 227–245, 2017.
- [49] A. Sinha and R. S. Shekhawat, "Olive Spot Disease Detection and Classification using Analysis of Leaf Image Textures," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 2328–2336, 2020, doi: [10.1016/j.procs.2020.03.285](https://doi.org/10.1016/j.procs.2020.03.285).
- [50] H. Alaa, K. Waleed, M. Samir, M. Tarek, H. Sobeah, and M. A. Salam, "An intelligent approach for detecting palm trees diseases using image processing and machine learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 7, pp. 434–441, 2020, doi: [10.14569/IJACSA.2020.0110757](https://doi.org/10.14569/IJACSA.2020.0110757).
- [51] S. Debnath et al., "Identifying individual nutrient deficiencies of grapevine leaves using hyperspectral imaging," *Remote Sens.*, vol. 13, no. 16, pp. 1–21, 2021, doi: [10.3390/rs13163317](https://doi.org/10.3390/rs13163317).
- [52] Z. Wang, M. Hu, and G. Zhai, "Application of deep learning architectures for accurate and rapid detection of internal mechanical damage of blueberry using hyperspectral transmittance data," *Sensors (Switzerland)*, vol. 18, no. 4, pp. 1–14, 2018, doi: [10.3390/s18041126](https://doi.org/10.3390/s18041126).
- [53] T. Islam, R. U. B. Rizan, Y. A. Tusher, M. Shafiuzzaman, M. A. Hossain, and S. Galib, "Nitrogen fertilizer recommendation for paddies through automating the Leaf Color Chart (LCC)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 745–752, 2020, doi: [10.14569/IJACSA.2020.0110891](https://doi.org/10.14569/IJACSA.2020.0110891).
- [54] M. V. Latte and S. Shidnal, "Multiple nutrient deficiency detection in paddy leaf images using color and pattern analysis," *Int. Conf. Commun. Signal Process. ICCSP 2016*, pp. 1247–1250, 2016, doi: [10.1109/ICCSP.2016.7754352](https://doi.org/10.1109/ICCSP.2016.7754352).
- [55] J. Drdsh, D. Eleyan, and A. Eleyan, "A Prediction Olive Diseases Using Machine Learning Models, Decision Tree and Naive Bayes Models," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 18, pp. 4231–4240, 2021.
- [56] S. S. Manhas, "Nutrient Deficiency Detection in Leaves using Deep Learning," 2021.
- [57] F. Jiang, Y. Lu, Y. Chen, D. Cai, and G. Li, "Image recognition of four rice leaf diseases based on deep learning and support vector machine," *Comput. Electron. Agric.*, vol. 179, no. April, p. 105824, 2020, doi: [10.1016/j.compag.2020.105824](https://doi.org/10.1016/j.compag.2020.105824).
- [58] Z. Chen and X. Wang, "Model for estimation of total nitrogen content in sandalwood leaves based on nonlinear mixed effects and dummy variables using multispectral images," *Chemom. Intell. Lab. Syst.*, vol. 195, no. July, p. 103874, 2019, doi: [10.1016/j.chemolab.2019.103874](https://doi.org/10.1016/j.chemolab.2019.103874).
- [59] V. Mande, "Estimation of chlorophyll based on FPGA and Matlab," 2017 *Int. Conf. Nascent Technol. Eng. ICNTE 2017 - Proc.*, pp. 1–5, 2017, doi: [10.1109/ICNTE.2017.7947986](https://doi.org/10.1109/ICNTE.2017.7947986).
- [60] R. Sathyavani, K. JaganMohan, and B. Kalaavathi, "Classification of nutrient deficiencies in rice crop using denseNet-BC," *Mater. Today Proc.*, no. xxxx, 2021, doi: [10.1016/j.matpr.2021.10.466](https://doi.org/10.1016/j.matpr.2021.10.466).

Automatic Fake News Detection based on Deep Learning, FastText and News Title

Youssef Taher¹

Center of Guidance and Planning
Rabat, Morocco

Adelmoutalib Moussaoui²

Faculty of Sciences and Technologies
Moulay Ismail University
Errachidia, Morocco

Fouad Moussaoui³

FST of Errachidia, Moulay Ismail
University, Errachidia, Morocco

Abstract—As a range of daily phenomena, Fake News is quickly becoming a longstanding issue affecting individuals, public and private sectors. This major challenge of the connected and modern world can cause many severe and real damages such as manipulating public opinion, damaging reputations, contributing to the loss in stock market value and representing many risks to the global health. With the fast spreading of online misinformation, checking manually Fake News becomes ineffective solution (not obvious, difficult and takes a long time). The improvement of Deep Learning Networks (DLN) can support with high degree of accuracy and efficiency the classical processes of Fake News spotting. One of the keys improvement strategies are optimizing the Word Embedding Layer (WEL) and finding relevant Fake News predicting features. In this context, and based on six DLN architectures, FastText process as WEL and Inverted Pyramid as News Articles Pattern (IPP), the present paper focuses on the assessment of the first news article feature that is hypothesized as affecting the performances of fake news predicting: News Title. By assessing the impact that the Embedding Vector Size (EVS), Window Size (WS) and Minimum Frequency of Words (MFW) in News Titles corpus can have on DLN, the experiments carried out in this paper showed that the News Title feature and FastText process can have a significant improvement on DLN fake news detection with accuracy rates exceeding 98%.

Keywords—Fake news; automatic detection; deep learning; FastText; news title

I. INTRODUCTION

Nowadays, fake news and sophisticated disinformation can have serious real world negative effects [1]. Often the main objectives of these false information is to intentionally deceive, gain attention, manipulate public opinion or to damage reputations.

During a time of uncertainty or crisis (ex. Covid-19), people are more likely to believe the false rumour they find on web, social media as well as online newspapers if it appeals to their emotions. As shown in Table I, this phenomenon of fake news can fall into different categories and take on different faces.

Deciphering these massive, instantaneous and heterogeneous daily news categories are valid or not becomes a serious challenge. In the era of social media, these false information and hoaxes spread freely, wider and more faster than ever before. These digital platforms enable novel forms

of communication, affect and accelerate the way individuals interpret daily developments.

Recently, in many democratic systems fake news distort and change how the electoral campaigns of candidates and political parties [2]. In the post-election period, significant number of websites and social media publish and share falsified or heavily biased information and stories, which calls into question the legitimacy of the elections.

In business and economics systems fake news and sophisticated disinformation are currently a hot topics. On the world economy, disinformation and hoaxes have direct and greater impacts. As example of consequences, every year fake news contribute to the loss in stock market value and investors can lose money due this problem [3].

In health systems, false information can spread easily and widely than pandemics, and can become accepted as true [4]. False information about virus can do real harm with often dangerous consequences. As examples, misleading information about medical services, medical products, treatments, official sources and guidelines can directly endangering the public health.

To develop the ability to decipher these fake news categories, various techniques and approaches have been developed. These solutions can be classified into two classes of methods (Fig. 1).

The first class is manual. This class is mainly based on comparing real news with unverified news by visiting fact checking sites.

TABLE I. FAKE NEWS CATEGORIES

Fake news categories	
▪	Blatantly false articles
▪	True articles with some false interpretations
▪	Pseudoscientific articles
▪	Opinion articles disguised as news
▪	Satirical articles
▪	Articles comprised of mostly text, tweets and quotes from other people

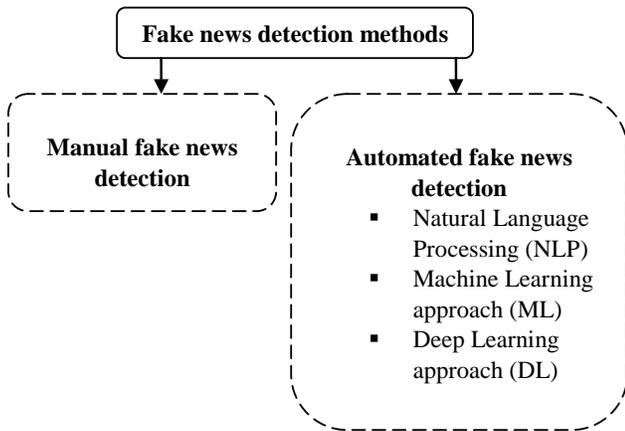


Fig. 1. Fake News Detection Approaches.

The second class is based on automated systems. To detect and predict misinformation, these recent systems exploit many important benefits of Natural Language Processing (NLP) and Artificial Intelligence (AI) [5] such as Machine Learning (ML) algorithms and Deep Learning (DL) networks (Singh et al., 2021)[6].

Recently, the improvement of these automated models can support with high degree of accuracy and efficiency the classical process of Fake News spotting. One of the key improvement strategies is optimizing the main Fake News predicting features and Word Embedding Layer (WEL).

In this context, and based on many DLN architectures and the news articles pattern IPP (Inverted Pyramid Pattern), the present paper focuses on the assessment of the first news feature that is hypothesized as affecting the performances of automatic Fake News spotting: News articles title.

The experiments carried out in this paper assess the impact of this key feature on Fake News DLN predicting performances by using six Deep Learning models: Simple LSTM, Stacked LSTM with two layers, Bidirectional LSTM, Simple GRU, Stacked GRU with two layers and Bidirectional GRU. These DLN models are feed by several rows of real and Fake News titles. Each title is a collection of English language words. To improve the embedding process of this corpus, the FastText process is used as a first embedding layer.

For each used model, the experiments assess the impact that Embedding Vector Size (EVS), Window Size (WS) and Minimum Frequency of Words (MFW) in the used corpus can have on Fake News spotting performances (execution time, loss and accuracy). To illustrate the diagnostic ability of each used model as its discrimination threshold is varied, the Receiver Operating Characteristics Curve (ROC) are plotted and the AUC values (Area Under the ROC Curve) are calculated by testing several values of EVS, WS and MFW.

Compared to many recent relevant studies on automatic fake detection based on the DLN architectures (summarized in paragraph III-B), the main objective of these experiments is to improve the performance of the execution time, loss and accuracy by testing a new embedding process (FastText), and reducing the dimensionality of the used articles news data by assessing the impact of the first news title feature that is

hypothesized as affecting the performances of automatic Fake News spotting.

This paper is organized as follows. Section 2 presents a brief overview of the used deep learning architectures. Section 3 provides a review of recent relevant studies on Fake Detection based on DLN and we summarize our findings. The contribution of this paper is presented in Section 4. This section presents the used architecture, embedding layer, used dataset, summarizes the five main features of the Inverted Pyramid Pattern (IPP), and discuss the impact of news title feature and FastText on fake news spotting performances. Finally, Section 5 presents our conclusions.

II. FUNDAMENTALS OF THE USED DEEP LEARNING ARCHITECTURES

In Deep Learning, Recurrent Neural Networks (RNN) are a family of Neural Networks [7]. These networks excels in learning by processing sequential data (one input follows another in time). RNN models use the current input and remember the preceding elements. The output at the current time step becomes the input to the next time step (Fig. 2).

Through its effectiveness, this kind of neural network is often used to handle text as news articles, tweets, comments, and have shown an important success in many Natural Language Processing (NLP) projects (Machine Translation, Speech Recognition, Generating Image Descriptions).

The basic architecture of RNN networks is Vanilla RNN (RNN network with single hidden layer). To deal with many limits of this RNN class, researchers have invented more advanced types of RNNs [8] such as Stacked RNNs, Bidirectional RNNs, Deep Bidirectional RNNs, Long Short Term Memory Networks (LSTM) [9], and Gated Recurrent Unit Networks (GRU).

The Stacked (Deep) RNN networks use multiple hidden RNN layers (Fig. 3). This architecture stacks multiple layers on top of each other. Each layer contains multiple cells, processes some part of the project tasks and passes it on to the next layer. The last layer provides the output. This approach (processing pipeline) has many potential benefits: exponentially more efficient to represent some functions and can for example extract more abstract features of news titles. However, these networks suffer from the vanishing gradient problem in the vertical direction.

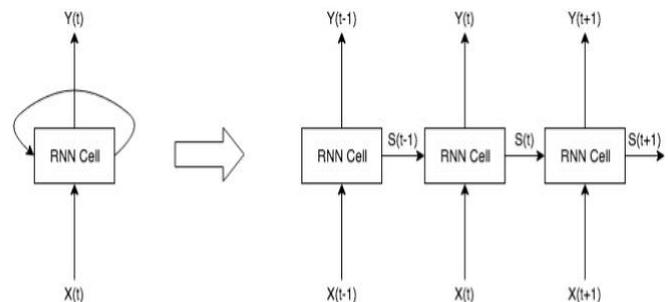


Fig. 2. Vanilla RNN.

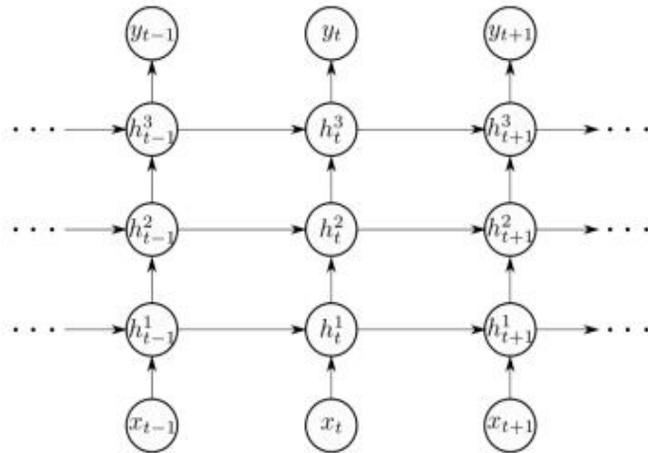


Fig. 3. Stacked RNN.

Bidirectional Recurrent Neural Networks (BRNNs) (Fig. 4) put two independent RNNs together without interacting with one another [10]. The first RNN network feeds the input sequence in normal time order (positive time direction). The second one feeds the input in reverse time order (negative time direction). Therefore, the model receives information from both past and future states. At each time, the output of BRNNs is computed after passing the merged results (by concatenating, adding, multiplying) of the forward and backward layers into the sigmoid function. Applied to fake news spotting process, this approach can improve the model performance and accuracy by obtaining the context in two directions compared to unidirectional RNN.

The Recurrent Neural Networks presented above suffer from vanishing and exploding gradient problems [11]. One of the important approaches to deal with these problems is Long Short Term Memory Networks (LSTM) [12]. To learn, remember and store relevant information for learning, these networks use a memory unit. By using a gating mechanism, LSTM architecture decides to pass the information to the next layer or forget the information it has. Fig. 5 summarizes the basic cell of LSTM network.

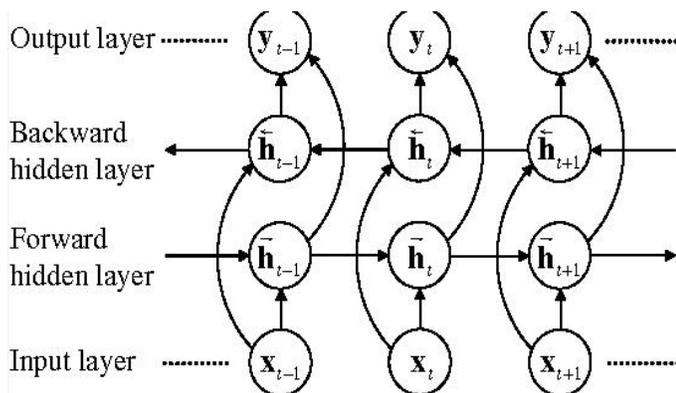


Fig. 4. Bidirectional RNN.

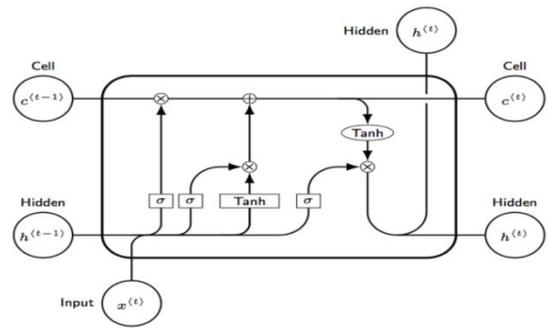


Fig. 5. Basic Cell of LSTM Network.

Each LSTM cell has three inputs ($h(t-1)$, $c(t-1)$, $x(t)$) and two outputs ($h(t)$, $c(t)$).

$x(t)$: the input at time t .

$h(t-1)$: the previous hidden state.

$h(t)$: the current hidden state.

$c(t-1)$: the previous cell state (the explicit memory unit).

$c(t)$: the current cell state.

The forget gate (first sigmoid layer with the inputs $x(t)$ and $h(t-1)$) selects the amount of information of the previous cell to be included. The input gate (second sigmoid layer with the inputs $x(t)$ and $h(t-1)$) decides what new information is to be added to the cell. These sigmoid layers determine the information to be stored in the cell state. By calculating the point-wise multiplication of the result of the tanh layer and the result of the input gate (Fig. 5), LSTM cell decides the amount of information to be added to the cell state. To produce $c(t)$, the result of this point-wise multiplication is added with the result of the first sigmoid layer multiplied by $c(t-1)$. By using a sigmoid and a tanh layer, the LSTM cell calculates the output. The sigmoid layer decides which part of $c(t)$ will be present in the output. The tanh layer shifts the output in the range of $[-1,1]$.

To simplify the internal design and to improve the design complexity, Cho proposed the Gated Recurrent Unit Network (GRU). This variant of the LSTM is based on two gates illustrated in Fig. 6: update gate (z) and a reset gate (r). To keep around how much previous memory, the GRU cell uses the update gate. To define how much information needs to be forgotten this cell uses the reset gate.

z : update gate

r : reset gate

Instead of the input, forget, and output gates in LSTM cell, GRU architecture uses this gated mechanism to capture dependencies of different time scales effectively, and retains LSTM's resistance to the vanishing gradient problem. The internal structure of GRU needs few computations to make updates to its hidden state.

accuracy over other methods namely CNN, vanilla RNN and unidirectional LSTM.

C. Used Embedding Process

Producing efficiently numerical dense vector (word embedding) is a key process for many news articles processing tasks such as articles news classification, fake news predicting, etc.

The optimized numerical vectors can encode efficiently the semantic information, measure the semantic similarity between two words in news articles, and use these numerical vectors as news articles features [20].

One of the most popular techniques used to create and to learn these vectors is Word2Vec [21]. This model developed by Google supports supervised learning and unsupervised learning. It based on two methods involving simple Neural Networks with one hidden layer: the Skip-Gram model and the Continuous Bag-of-Words model (CBOW). The word vectors are learned via backpropagation and stochastic gradient descent. The Skip-Gram model can use the target word to predict the context. The CBOW method takes the context of each word as the input and tries to predict the word corresponding to the context.

As a modified version (extension) of Word2Vec (Skip-Gram and CBOW) presented above, the FastText process [22] is used as embedding layer in the proposed architecture.

FastText is a library for efficient word embeddings and text classification. This library is developed by the Facebook research team has shown excellent results on many Natural Language Processing (NLP) projects (faster with superior performance).

To improve the used vector representations, FastText process treats each word in news titles corpus as composed of character n-grams (split each word in multiple n parts). It may be bigram, trigram, etc. (Table II). The character n-grams of length n can be generated by sliding a window of n-characters from the start till the end.

By providing Skip-Gram and CBOW models, FastText process computes News words representations. Each word of the used dictionary is represented by the sum of the vector representations of its n-gram. Consequently, and by averaging the vectorized representation of all its constituent n-grams, the embedding process can generate News Titles word vectors for the words that does not appear in the training corpus.

In addition to this last benefit, this used embedding layer is significantly better than the original Word2Vec on syntactic tasks, especially in the case of training corpus with small sizes (case of the present study).

The pooling strategy of FastText can generate a huge number N of unique n-grams. To bind the memory requirements, a hashing process is used. Each character n-gram is hashed to an integer between 1 to H (bucket size). Therefore, FastText learns a total H embeddings instead of learning a total N embedding (Fig. 7).

TABLE II. EXAMPLES OF DIFFERENT LENGTH CHARACTER N-GRAMS

Word	Length(n)	Character n-grams
author	3	<au, aut, uth, tho, hor, or>
author	4	<aut, auth, utho, thor, hor>
author	5	<auth, autho, uthor, thor>
author	6	<autho, author, uthor>

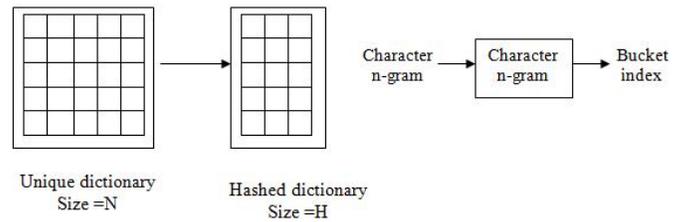


Fig. 7. Hashed Dictionary.

D. Used News Article Pattern

Usually, news articles describe events, persons, occurrences, experiences, places and other topics by following a particular pattern (how information should be prioritised and structured) [23].

One of the most common used patterns is the Inverted Pyramid (IP) which often composed of five important news features (Fig. 8).

News information in the IP pattern is presented in descending order of importance. The first part (feature) is the title (headline). This feature tells what the news is about. The second feature shows who wrote the news (byline). The third feature is the lead (first paragraph). This paragraph summarizes the main and important facts of the news, and based on 5 W's (who, what, when, where, and why) and how. The fourth feature is the body. This part is the core information and details about the news, which supports and amplifies the lead. The fifth feature is the ending which usually gives something to think about.

Often the order of IP pattern allows reading quickly the most crucial information, and estimating an initial manual spotting of Real or Fake News.

Based on IP order of importance, this paper assesses the impact of news title feature (first feature) that is hypothesized as affecting the performances of fake news spotting.

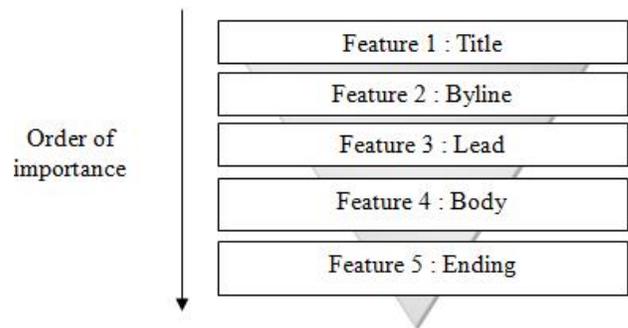


Fig. 8. Used Article News Pattern IP.

E. Used Dataset

The present study uses two Fake and Real News datasets from Kaggle source [24]. This news dataset is based on two 2 files (two lists of news articles): Fake News and True News files. Each file has four features: News Title, News Lead, News Subject and News Date (Table III).

TABLE III. FEATURES OF THE USED DATASET

News title	News lead	News subject	News date
The title of the article	The lead of the article	The subject of the article	The date at which the article was posted

After downloading, these two files are merged into a single dataset. This news dataset is labelled by adding a new feature called News Label. "0" value is assigned to Fake News (FN) and "1" value is assigned to True News (TN).

Fig. 9 summarizes the number of each label value. News Title data are extracted to fed FastText embedding layer and building the used six DLN models.

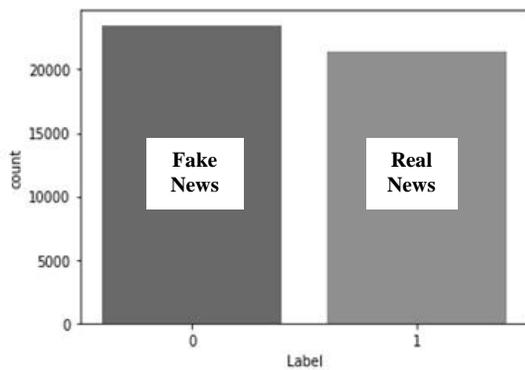


Fig. 9. Number of used Fake and True News Articles.

The encoded news data are tokenized, created and padded by using TensorFlow [25] and Keras [26] preprocessing tools (Fig. 10).

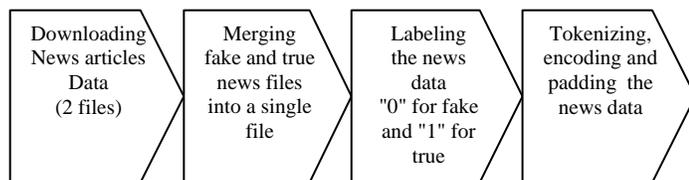


Fig. 10. Articles News Data Pre-Processing Process.

F. Automatic Fake News Detection based on DLN, FastText and News Titles

1) Assessment of execution time performance: Usually fake news predicting projects focus on the accuracy of Deep Learning (DL) or Machine Learning (ML) models that they are using. However, optimizing DLN execution time is one of the important processes that can improve the performances of Fake News detection software. Particularly, when DL model is used as a specific service or part of service installed on personal computers or other devices with limited resources [27]. In this context, the first experiments of this study assess

the time performances of the used embedding layer (FastText process) based on news title feature.

Generally, the duration of the embedding process can be due to the hardware platform architecture (Central Processing Unit (CPU), Graphics Processing Unit (GPU), Random Access Memory (RAM)), internal or external interruptions during the computation and the used libraries if are optimized or not.

The first experiments used the hardware configuration summarized in the following Table IV:

TABLE IV. USED HARDWARE CONFIGURATION

Processor	Intel (R) Core (TM) i7- 4610 M
Processor's speed	CPU@3.00 GHz
Memory size	8.00 Go

The main objective of these first experiments is estimating and assessing the impact that the main parameters of the embedding layer (Embedding Vector size (EVS), Window Size (WS), Minimum Frequency of Words (MFW)) can have on time performance. These experiments start by feeding the embedding layer by several EVS values and setting the used maximum dimension to 140 (≈ half of the maximum length of News Titles). As shown in Fig. 11, the execution time of the embedding layer increases with important rate by increasing EVS values. The average execution time obtained by smaller sizes (less than 40) is relatively small compared to longer embedding vectors (greater than 100). If the embedding size go from 10 to 100, the execution time will double its value.

The second experiment assesses the effect of WS parameter (window of surrounding context words) on the execution time. We fed the embedding layer by several WS sizes on a scale of 1 to 10. As shown in Fig. 12, as WS values increase, the execution time increases. When WS changes from WS =1 to WS = 10, the approximate change in the execution time is around 67 percent (67, 01%). Large values of WS (greater than WS = 5) increase slightly the execution time (≈3 seconds of difference from WS = 6 to WS = 10).

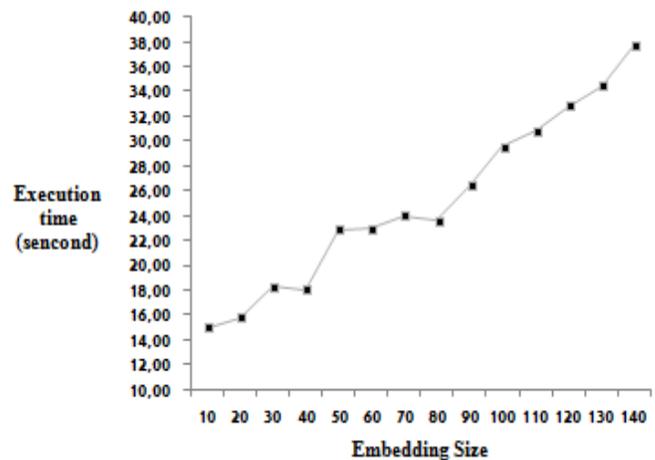


Fig. 11. Variation of the Execution Time Values with Different Embedding Vector Sizes. WS =2, MFW=2.

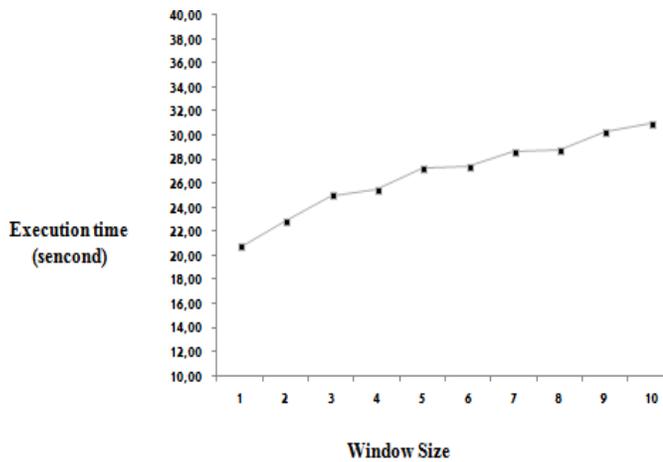


Fig. 12. Variation of the Execution Time Values with Different Window Sizes. EVS=50, MFW=2.

In contrast with the impact of the parameters EVS and WS discussed above, the execution time of the embedding layer decreases by increasing the MFW values in the used corpus. As shown in Fig. 13, the execution time can be reduced to half ($\approx 55,96\%$) of its initial value when MFW changes from MFW =1 to MFW = 10.

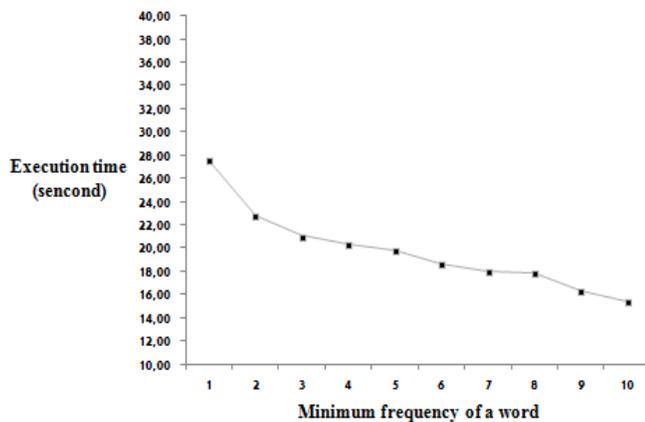


Fig. 13. Variation of the Execution Time Values with Different MFW Values. EVS=50, WS=2.

As a main result of these first experiments, the execution time of the used embedding layer can be improved if the set of parameters (WS, MFW, EVS) is optimized. Consequently,

TABLE V. VARIATION OF LOSS VALUES WITH DIFFERENT DEEP LEARNING MODELS AND EVS VALUES. WS =2, MFW=2

Models	Embedding Vector Size (EVS)				
	10	50	80	100	140
Simple GRU	1,16%	0,62%	0,53%	0,52%	0,49%
Bidirectional GRU	1,12%	0,46%	0,31%	0,27%	0,26%
Stacked GRU	1,11%	0,55%	0,55%	0,55%	0,48%
Simple LSTM	1,09%	0,63%	0,59%	0,55%	0,55%
Stacked LSTM	1,05%	0,57%	0,54%	0,53%	0,51%
Bidirectional LSTM	0,96%	0,44%	0,30%	0,27%	0,25%

this optimized execution time can positively impact the total duration of DLN computation process (embedding layer and DLN models execution times).

2) *The impact of embedding vector size (EVS) on fake news detection performances:* This experiment assesses the effects of FastText process and News title feature on Fake News detection performances by building the used six DLN architectures with less EVS sizes and then more EVS sizes. The used EVS ranges from 10 to 140 (140 \approx half of the maximum length of news titles). As shown in Table V, as EVS values increase, the loss values decrease slightly for all the used six DLN architectures. When EVS changes from EVS =10 (small size) to EVS = 140 (high-dimensionality), the approximate decreases in the loss are around 0.67% for simple GRU, 0.86% for bidirectional GRU, 0.63% for stacked GRU, 0.54% for simple LSTM, 0.54% for stacked LSTM, and 0.71% for bidirectional LSTM. According to the Table VI, the News Title feature has a significant influence on the accuracy for all used detection models. These six architectures can achieve an accuracy that exceeds 98% for small and high values of EVS. The accuracy value increases slightly by increasing EVS. If EVS go from 10 to 140, the approximate increases in accuracy are around 0.83% for simple GRU, 1,06% for bidirectional GRU, 0,76% for stacked GRU, 0,72% for simple LSTM, 0,71% for stacked LSTM, and 0,99% for bidirectional LSTM. Compared to the other used models, the bidirectional LSTM network achieved the best accuracy for all used EVS values.

The ROC curves plotted with different models and EVS values show that News Title feature and the six used architectures provide a high fake news detection accuracy (Fig. 14). These curves show that there is significant improvement in News classification accuracy with lower decision thresholds, in particular with the bidirectional LSTM network. The AUC (Area Under ROC Curve) ranges in value from 0.95 to 0.97 when the used EVS values range from 10 to 140.

As a main result of this experiment, it's possible to achieve better Fake News detection performance by choosing a low Embedding Vector Sizes (EVS).

TABLE VI. VARIATION OF ACCURACY VALUES WITH DIFFERENT DEEP LEARNING MODELS AND EVS VALUES. WS =2, MFW=2

Models	Embedding Vector Size (EVS)				
	10	50	80	100	140
Simple GRU	98,63%	99,30%	99,45%	99,45%	99,46%
Bidirectional GRU	98,68%	99,50%	99,69%	99,74%	99,74%
Simple LSTM	98,68%	99,31%	99,37%	99,40%	99,40%
Stacked GRU	98,74%	99,29%	99,33%	99,42%	99,50%
Stacked LSTM	98,75%	99,38%	99,44%	99,44%	99,46%
Bidirectional LSTM	98,82%	99,61%	99,73%	99,79%	99,81%

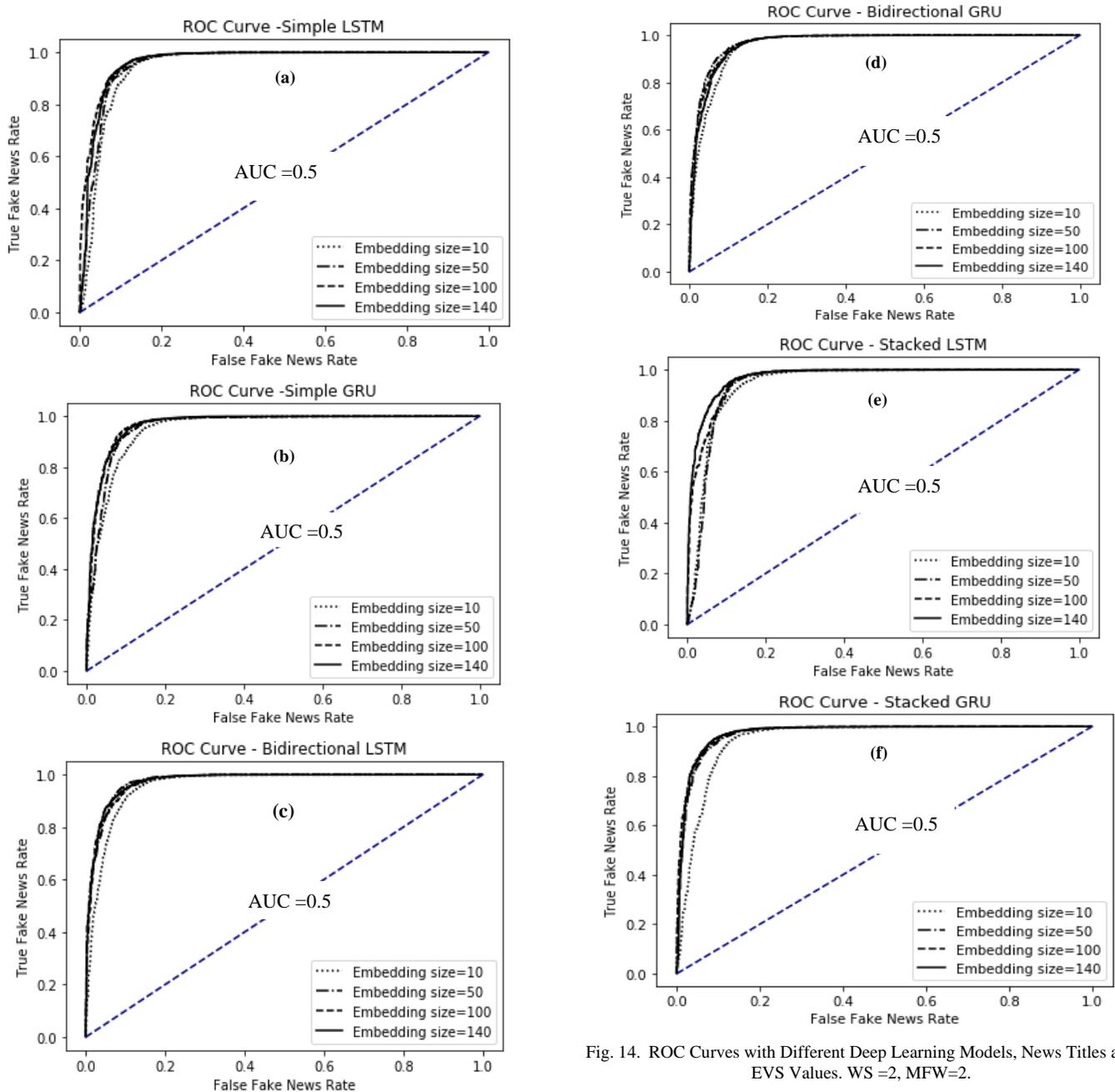


Fig. 14. ROC Curves with Different Deep Learning Models, News Titles and EVS Values. WS =2, MFW=2.

TABLE VII. VARIATION OF LOSS VALUES WITH DIFFERENT DEEP LEARNING MODELS AND WS VALUES. EVS =50, MFW=2

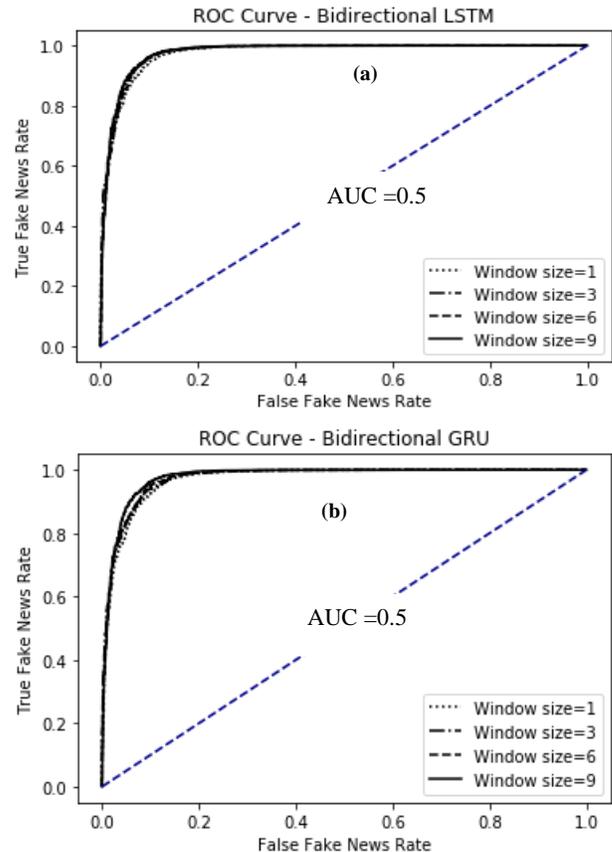
Window Size (WS)						
	1	3	5	7	9	10
Bidirectional LSTM	0,35%	0,28%	0,30%	0,31%	0,32%	0,29%
Bidirectional GRU	0,36%	0,29%	0,30%	0,31%	0,34%	0,31%
Simple GRU	0,61%	1,28%	0,84%	0,76%	0,64%	0,30%
Stacked GRU	0,62%	0,64%	0,68%	0,53%	0,65%	0,60%
Simple LSTM	0,62%	0,51%	0,55%	0,50%	0,52%	0,54%
Stacked LSTM	0,64%	0,55%	0,58%	0,51%	0,55%	0,57%

TABLE VIII. VARIATION OF ACCURACY VALUES WITH DIFFERENT DEEP LEARNING MODELS AND WS VALUES. EVS =50, MFW=2

Window Size (WS)						
	1	3	5	7	9	10
Bidirectional LSTM	99,32%	99,65%	99,01%	99,12%	99,27%	99,68%
Bidirectional GRU	99,61%	99,70%	99,69%	99,66%	99,63%	99,68%
Simple GRU	99,32%	99,29%	99,24%	99,41%	99,29%	99,31%
Stacked GRU	99,31%	99,44%	99,40%	99,44%	99,42%	99,41%
Simple LSTM	99,28%	99,39%	99,35%	99,45%	99,40%	99,36%
Stacked LSTM	99,61%	99,73%	99,67%	99,68%	99,67%	99,71%

3) *The impact of window size (WS) on fake news detection performances:* Generally, Window Size (WS) has the impact of giving more importance to closer words. Smaller WS lead to similar interchangeable words. Larger WS lead to similar related words. This experiment assesses the impact of WS by using less WS values and then more WS values. The used WS ranges from WS=1 to WS=10. According to the Table VII, increasing WS values causes a small variance of loss values. When WS changes from WS =1 (small Window Size) to WS = 10 (high Window Size), the difference between the minimum and the maximum of loss values does not exceed 0,07% for Bidirectional LSTM and Bidirectional GRU, 0,15% for simple LSTM, Stacked LSTM / Stacked GRU and 1% for simple GRU.

These architectures achieve high accuracy exceeds 98% for small and high used values of WS. If WS go from 1 to 10, the approximate difference between the minimum and the maximum of accuracy values does not exceed 0,09% for Bidirectional GRU, 0,17% for simple LSTM / simple GRU, 0,13% for Stacked LSTM / Stacked GRU and 0.67% for Bidirectional LSTM (Table VIII). By testing smaller and larger WS values, the ROC curves plotted with different Window Sizes show that news title feature and FastText process provide a high Fake News classification accuracy (Fig. 15).



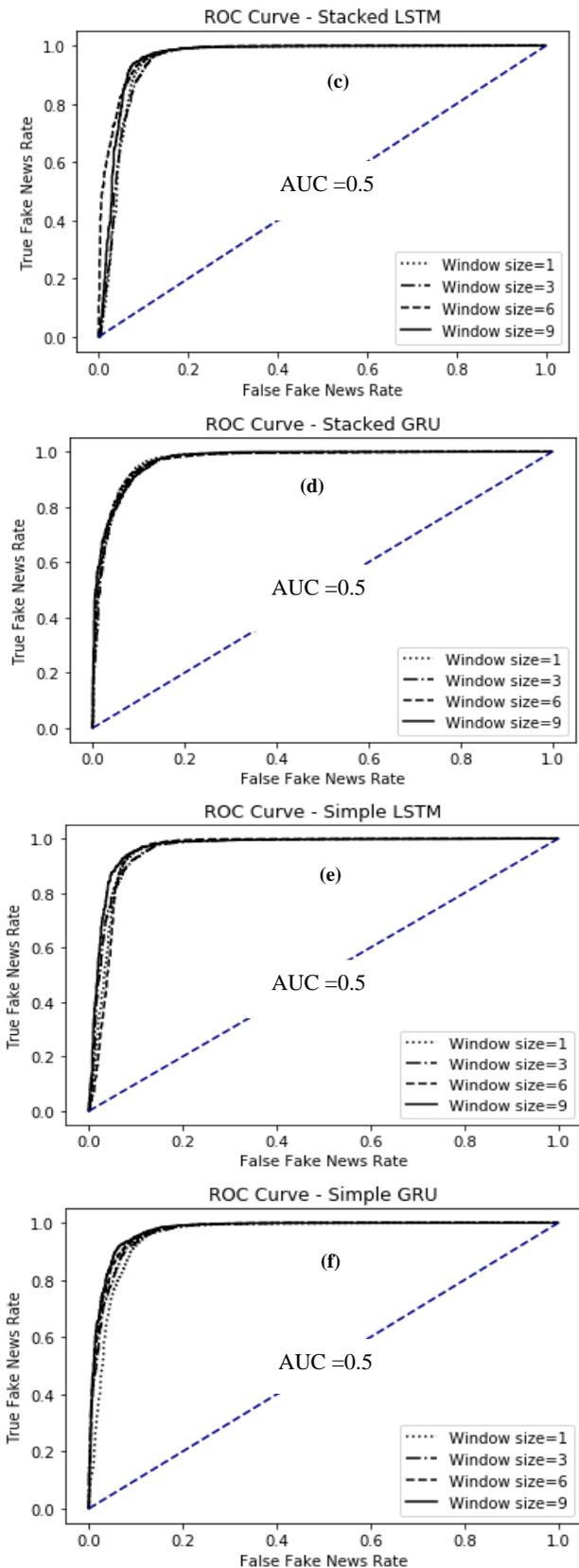


Fig. 15. ROC Curves with Different Deep Learning Models and WS Values. EVS =50, MFW=2.

Increasing WS values impacts slightly the ROC decision thresholds. Important improvement in Fake News classification is obtained from lower decision thresholds for all the six used models. When WS values range from 1 to 10, AUC ranges in value from 0.96 to 0.97 for LSTM architectures and from 0.96 to 0.98 for GRU architectures. As main result of these experiments, the use of News Title feature with low dimensions of Windows Size (WS between two and three) are enough to capture enough information with to detect Fake News with high performance.

4) *The impact of minimum frequency of a word (MFW) on fake news detection performances:* The hyper-parameter MFW (Minimum Frequency of Word) specifies the minimum frequency of a word in the News Titles corpus for which the word embedding will be generated. This last experiment assesses the impact of MFW by testing less MFW values and then more MFW values. The used MFW ranges from 1 to 9. As shown in the Table IX, the loss value is slightly decreased by increasing the values of MFW. If MFW go from 1 to 9, it decreases by 0,11% for Bidirectional LSTM, 0,16% for Bidirectional GRU, 0,04% for Stacked LSTM, 0,17% for Stacked GRU, 0,09% for Simple LSTM and 0,11% for Simple GRU. The minimum of loss values is observed for high values of MFW (MFW = 5 for simple GRU, MFW = 9 for all other models). The stacked GRU model shows the higher loss value for MFW between 1 and 5.

Compared to the other used models, Bidirectional LSTM and GRU models show the higher accuracy for all used MFW values. The ROC curves plotted with different MFW values show that news title feature used as input of FastText embedding layer provides a high Fake News classification accuracy for smaller and larger MFW values (Fig. 16). Increasing MFW value impacts slightly the ROC decision thresholds. Important improvement in the Fake News classification is obtained from lower decision thresholds for all the used models.

For higher values of MFW (MFW between 7 and 9), this maximum value is obtained by the simple GRU model.

The six used architectures achieve high accuracy exceeds 98% for small and high used values of MFW. According to the Table X, news title feature has a significant influence on the accuracy for all the used MFW values. As MFW values increase, the accuracy values increase slightly for all the used models. When MFW changes from MFW =1 (small frequency) to MFW = 9 (high frequency), the approximate increases in accuracy are around 0.14% for Bidirectional LSTM, 0.17% for Bidirectional GRU, 0.12% for Stacked / simple LSTM, 0.16% for Simple GRU, and 0.25% for Stacked GRU.

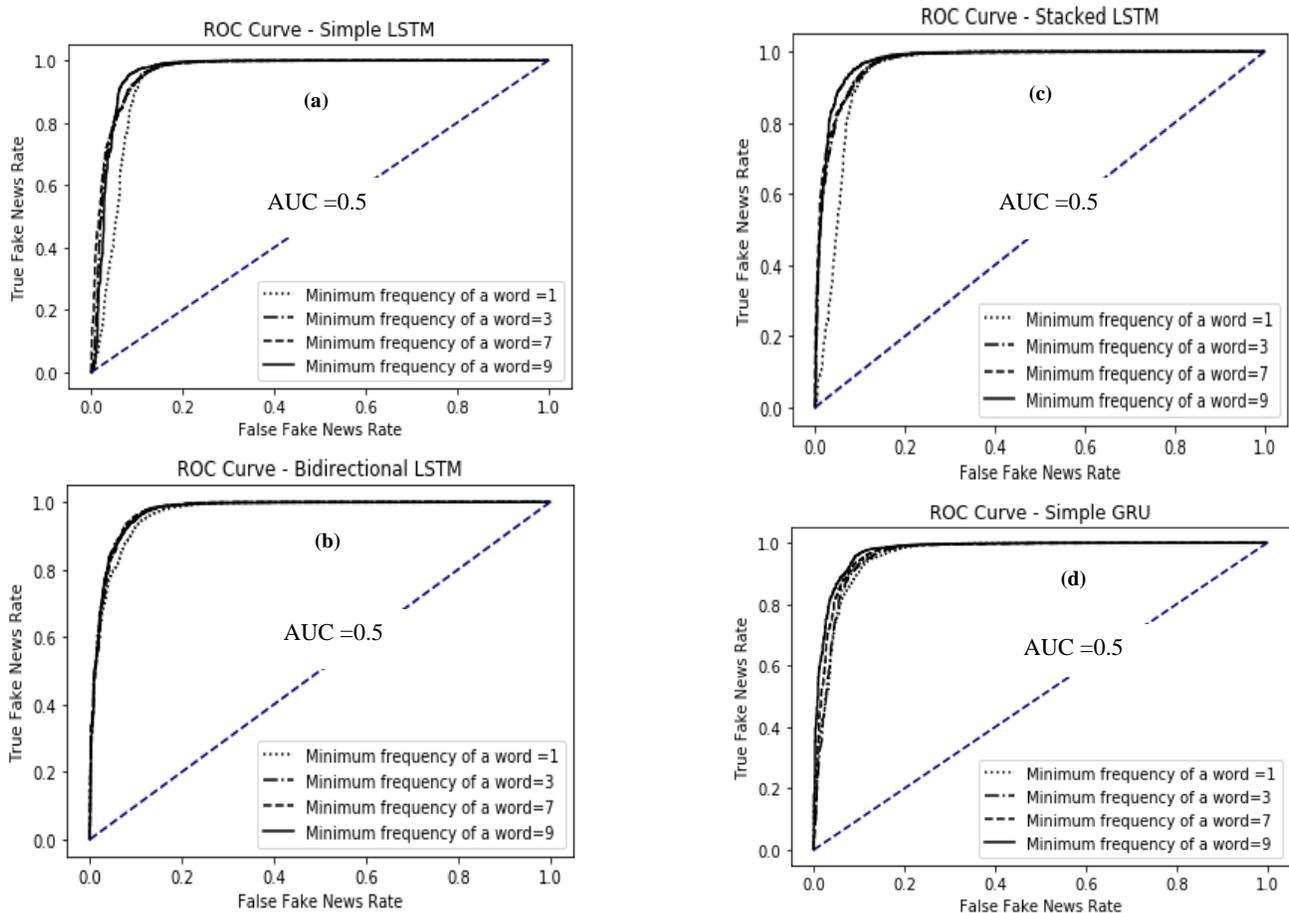
When MFW values range from 1 to 9, the AUC (Area Under the ROC Curve) ranges in value from 0.96 to 0.97 for all used LSTM architectures and from 0.96 to 0.98 for all used GRU architectures. The findings of this last experiment showed that News titles with low size of the minimum frequency of words (MFW) didn't affect significantly the performances of Fake News detection.

TABLE IX. VARIATION OF LOSS VALUES WITH DIFFERENT DEEP LEARNING MODELS AND MFW VALUES. EVS =50, WS=2

Models	Minimum frequency of a word (MFW)				
	1	3	5	7	9
Bidirectional LSTM	0,46%	0,39%	0,39%	0,39%	0,35%
Bidirectional GRU	0,47%	0,44%	0,42%	0,39%	0,31%
Stacked LSTM	0,60%	0,60%	0,58%	0,56%	0,56%
Simple LSTM	0,65%	0,59%	0,58%	0,58%	0,56%
Simple GRU	0,72%	0,66%	0,58%	0,60%	0,61%
Stacked GRU	0,75%	0,67%	0,64%	0,58%	0,58%

TABLE X. VARIATION OF LOSS VALUES WITH DIFFERENT DEEP LEARNING MODELS AND MFW VALUES. EVS =50, WS=2

Models	Minimum frequency of a word (MFW)				
	1	3	5	7	9
Bidirectional LSTM	99,56%	99,62%	99,65%	99,70%	99,70%
Bidirectional GRU	99,51%	99,55%	99,56%	99,59%	99,68%
Stacked LSTM	99,26%	99,32%	99,35%	99,39%	99,38%
Simple LSTM	99,29%	99,33%	99,35%	99,38%	99,41%
Simple GRU	99,20%	99,23%	99,33%	99,35%	99,36%
Stacked GRU	99,13%	99,25%	99,30%	99,37%	99,38%



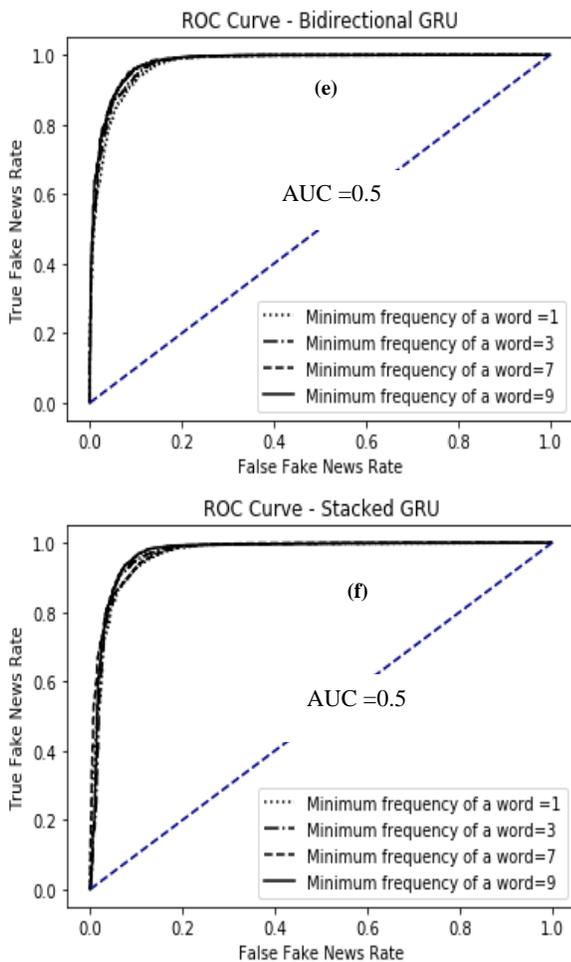


Fig. 16. ROC Curves with Different Deep Learning Models and MFW Values. EVS =50, WS=2.

IV. CONCLUSION

Today, Fake news phenomena are one of the world's top global risks. Therefore, social, political and economic environments need to be well-equipped to decipher and detect this massive, instantaneous and heterogeneous daily misinformation.

One of the promising fields of automated systems used to deal with this problem is Artificial Intelligence (AI), especially Deep Learning Networks (DLN).

In this context, the present paper focuses on the improvement of DLN Fake News detection by using Inverted Pyramid Pattern (IPP), and integrating the FastText process in many DLN architectures: Simple LSTM, Stacked LSTM, Bidirectional LSTM, Simple GRU, Stacked GRU and Bidirectional GRU. More precisely, this empirical study focuses on how news title feature and FastText embedding process can impact and improve the existing automatic DLN Fake News detection.

By testing these six architectures with many ranges of embedding layer main hyper-parameters (EVS, WS, MFW), these experiments showed that the news title feature and FastText process can have a significant improvement on DLN

automatic Fake News detection with accuracy rates exceeding 98%.

This paper is the first step of our future Fake News framework project based on many DLN architectures.

ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to Khalid AHAJI, Hassan ELHANI, Fouad MOUSSAOUI, Abdelmoutalib MOUSSAOUI who encouraged and helped us in doing a lot of research.

Any attempt at any level can't be satisfactorily completed without the support and guidance of my parents, sisters and friends.

REFERENCES

- [1] Á. Figueira and L. Oliveira, "The current state of fake news: Challenges and opportunities," *Procedia Computer Science*, vol. 121, pp. 817–825, 2017.
- [2] C. Sindermann, A. Cooper, and C. Montag, "A short review on susceptibility to falling for fake political news," *Current Opinion in Psychology*, vol. 36, pp. 44–48, 2020.
- [3] D. D. Parsons, "The impact of fake news on company value: Evidence from Tesla and Galena Biopharma," [Online]. Available: https://trace.tennessee.edu/utk_chanhonoproj/2328/. [Accessed: 28-Nov-2021].
- [4] M. A. Rivera, "Fake news and hospitality research," *International Journal of Hospitality Management*, vol. 85, p. 102473, 2020.
- [5] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," *Physica A: Statistical Mechanics and its Applications*, vol. 540, p. 123174, 2020.
- [6] D. S and B. Chitturi, "Deep neural approach to fake-news identification," *Procedia Computer Science*, vol. 167, pp. 2236–2243, 2020.
- [7] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.
- [8] Y. Ma and J. C. Principe, "A taxonomy for Neural Memory Networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 6, pp. 1780–1793, 2020.
- [9] Deep learning crash course for beginners with python: Theory and applications of artificial neural networks, CNN, RNN, LSTM and autoencoders using tensorflow 2: Contains exercises with solutions and hands-on projects. AI Publishing, 2020.
- [10] P. Bahad, P. Saxena, and R. Kamal, "Fake news detection using bi-directional LSTM-recurrent neural network," *Procedia Computer Science*, vol. 165, pp. 74–82, 2019.
- [11] S. Al-Abri, T. X. Lin, M. Tao, and F. Zhang, "A derivative-free optimization method with application to functions with exploding and vanishing gradients," *IEEE Control Systems Letters*, vol. 5, no. 2, pp. 587–592, 2021.
- [12] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [13] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali, "Diagnostic methods 2: Receiver operating characteristic (ROC) curves," *Kidney International*, vol. 76, no. 3, pp. 252–256, 2009.
- [14] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.
- [15] K. Englmeier, "The role of text mining in mitigating the threats from fake news and misinformation in times of Corona," *Procedia Computer Science*, vol. 181, pp. 149–156, 2021.
- [16] S. R. Sahoo and B. B. Gupta, "Multiple features based approach for automatic fake news detection on social networks using Deep Learning," *Applied Soft Computing*, vol. 100, p. 106983, 2021.

- [17] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet – a deep convolutional neural network for fake news detection," *Cognitive Systems Research*, vol. 61, pp. 32–44, 2020.
- [18] D. S. and B. Chitturi, "Deep neural approach to fake-news identification," *Procedia Computer Science*, vol. 167, pp. 2236–2243, 2020.
- [19] M. H. Goldani, S. Momtazi, and R. Safabakhsh, "Detecting fake news with Capsule Neural Networks," *Applied Soft Computing*, vol. 101, p. 106991, 2021.
- [20] M. Farouk, "Measuring text similarity based on structure and word embedding," *Cognitive Systems Research*, vol. 63, pp. 1–10, 2020.
- [21] D. Jatnika, M. A. Bijaksana, and A. A. Suryani, "Word2Vec model analysis for semantic similarities in English words," *Procedia Computer Science*, vol. 157, pp. 160–167, 2019.
- [22] J. Choi and S.-W. Lee, "Improving FastText with inverse document frequency of subwords," *Pattern Recognition Letters*, vol. 133, pp. 165–172, 2020.
- [23] A. Toxboe, "Article list design pattern," *UI Patterns*. [Online]. Available: <http://ui-patterns.com/patterns/ArticleList>. [Accessed: 30-Nov-2021].
- [24] C. Bisailon, "Fake and real news dataset," *Kaggle*, 26-Mar-2020. [Online]. Available: <https://www.kaggle.com/clmentbisailon/fake-and-real-news-dataset>. [Accessed: 30-Nov-2021].
- [25] P. S. Janardhanan, "Project Repositories for Machine Learning with tensorflow," *Procedia Computer Science*, vol. 171, pp. 188–196, 2020.
- [26] V.-H. Nhu, N.-D. Hoang, H. Nguyen, P. T. Ngo, T. Thanh Bui, P. V. Hoa, P. Samui, and D. Tien Bui, "Effectiveness assessment of keras based deep learning with different robust optimization algorithms for shallow landslide susceptibility mapping at tropical area," *CATENA*, vol. 188, p. 104458, 2020.
- [27] Y. Kwon, S. Lee, H. Yi, D. Kwon, S. Yang, B.-gon Chun, L. Huang, P. Maniatis, M. Naik, and Y. Paek, "Mantis: Efficient predictions of execution time, energy usage, memory usage and network usage on smart mobile devices," *IEEE Transactions on Mobile Computing*, vol. 14, no. 10, pp. 2059–2072, 2015.

Determine the Level of Concentration of Students in Real Time from their Facial Expressions

Bouhlal Meriem*, Habib Benlahmar, Mohamed Amine Naji, Elfilali Sanaa, Kaiss Wijdane

Laboratory of Information Technology and Modeling
Hassan II University of Casablanca, Faculty of Sciences, Casablanca, Morocco

Abstract—In teaching environments, student facial expressions are a clue to the traditional classroom teacher in gauging students' level of concentration in the course. With the rapid development of information technology, e-learning will take off because students can learn anytime, anywhere and anytime they feel comfortable. And this gives the possibility of self-learning. Analyzing student concentration can help improve the learning process. When the student is working alone on a computer in an e-learning environment, this task is particularly challenging to accomplish. Due to the distance between the teacher and the students, face-to-face communication is not possible in an e-learning environment. It is proposed in this article to use transfer learning and data augmentation techniques to determine the concentration level of learners from their facial expressions in real time. We found that expressed emotions correlate with students' concentration, and we designed three distinct levels of concentration (highly concentrated, nominally concentrated, and not at all concentrated).

Keywords—*Emotion recognition; level of concentration; transfer learning; data augmentation*

I. INTRODUCTION

In recent years, E-learning is very popular because this type of learning system uses modern educational technologies to implement an ideal learning environment by integrating information technology into the program.

E-learning is gaining prominence in universities, colleges, and industries by examining its advantages over traditional approaches, as students can access all the data they need for their research. Through webinars, they can access information they might not otherwise be able to access in person due to finances, distances, or time constraints. Depending on the level of understanding of the student, they can study at their own pace, which may increase their satisfaction with the course and reduce their stress levels. It also helps students with special needs or having difficulty getting to school for one reason or another, such as illness or unforeseen accidents.

In traditional classroom teaching, the experienced teacher can know the level of concentration by the students from their behaviors which can be studied by heart rate, pose, gesture, gaze, height of the voice and facial expressions and, accordingly, it can adjust the pace, improve the educational system and provide content. While the online environment separates teachers from students and students from students, there is a lack of face-to-face communication to understand students' emotions and cognitive state.

An effective individualized learning system must therefore be not only intelligent but also emotional. Researchers in neuroscience and psychology have found that emotions are largely related to cognition / concentration. These eLearning system models place particular emphasis on assessing learner emotional states and adjusting teaching strategies.

Among the challenges teachers face is examining how learners acquire course content, as noted in [1]. In order to improve the educational system, it is of paramount importance to ensure student concentration through participation in the learning environment.

Virtual classrooms were introduced as early as the mid-1990s [2]. Deconcentration of students is an issue that is addressed daily.

The way that students are taught is also a factor behind student concentration. Bradbury [3] reported that between 25% and 60% of students became bored in the classroom and lost concentration for a long time [4]. According to Ekman, Friesen, and Ellsworth, [5], facial expression can be the fastest way to understand an individual's emotions. You can use a student's emotional state during their learning period (in the classroom or another setting) to determine if they are paying attention to the content.

To identify students' concentration, other authors suggest using pupil dilation, which occurs when students view images of emotional arousal [6], or the length of time the eyes are closed [1]. Students' facial expressions can be captured by using the embedded webcam in their laptop computers in a typical e-learning environment. The teacher can use this information to determine the level of concentration of the students, and measure how focused they are (or aren't). The teacher can use this information to make the learning environment cost-effective.

The concentration of learners is influenced by several factors. The emotional life of learners has a profound impact on academic success, learning techniques, and academic achievement [7]. If emotions could be recognized as impacts on motivation and concentration, educational outcomes could be improved [8]. Emotion in this context corresponds to the psychological and physiological characteristics of a being, which are individual, efficient, and personal in nature, which relates to habits, manners, thoughts, and sensations [9]. It has been shown that facial muscles move in relation to different emotions, such as happiness, sadness, anger, fear, surprise, and disgust [10].

*Corresponding Author.

Most facial expression recognition approaches are based on a posed expression database, such as the Japanese Female Facial Expression Database (JAFFE), Cohn-Kanade Database (CK) which are built on six Universal emotions such as happiness, sadness, surprise, anger, disgust and fear due to the lack of database of facial expressions in educational environments.

The proposed system automatically identifies the emotional state of the learner based on facial expressions to know the level of concentration. This helps the teacher to identify slow learners who have difficulty understanding the lesson, therefore the lesson may be changed.

Students' engagement and concentration during learning is a prerequisite for successful learning effects and is positively correlated with students' academic success and development of higher-level abilities (Pascarella, Seifert, & Blauch, 2010). Effective detection of students' learning situations can provide information to instructors so that they can identify struggling students in real time. And this is so that a student's level of concentration and state of learning engagement can help intelligent tutoring systems provide students with individualized learning resources.

The purpose of this article is to examine the concentration level of students in a typical e-learning scenario, by analyzing their facial emotions in real time. In our analysis, we attempt to define how emotions affect concentration level and to devise a concentration index based on this. A facial emotion index is generated in real time by Python and the Keras algorithm, which is derived from the Haar-cascade algorithm and we will compare the transfer learning (a pre-trained convolutional neural network (CNN): VGG16, VGG19, XCEPTION, ALEXNET) models to determine the best performing model for our proposal.

This paper is structured as follows: Section 2 presents a review of related work; Section 3 discusses the methodology adopted for facial emotion recognition. The implementation environment, datasets, as well as the algorithms used, the experiment performed and the results are presented in Section 4. We concluded the paper by summarizing the research performed and providing some remarks in Section 5.

II. RELATED WORK

Currently available e-learning systems have a flaw: they do not monitor student concentration. Recent years have seen increased interest in finding clues to determine student concentration. Due to the absence of a teacher and the inability to grasp emotions, emotions, etc., are particularly important for students using standalone e-learning systems. According to the study conducted by Du, Tao, and Martinez [11], there are 22 different types of emotions: seven basic emotions and twelve compound emotions. There are forty-six fundamental muscles that form the UA of the face, including those that produce facial expressions. Based on individual UA, the system classifies facial categories by combining each individual UA (Fig. 1). In the case of AU12 and 25, if the system recognizes that the image indicates a "happy" emotion (Table I), the system will classify the image accordingly (Table I).

Bidwell and Fuchs [13] used an automated gaze system to measure student engagement. Classifiers were created based on video recordings of classrooms. A face tracking system was used to track student gaze. After the automated gaze model and observations of a panel of experts were combined, a Hidden Markov Model (HMM) was constructed. HMM incorrectly categorized the data, suggesting eight discrete categories of behaviors, but they were only able to determine whether students were "engaged" or "not engaged".

TABLE I. THE UNITED STATES OBSERVED UAS IN BOTH BASIC AND COMPOUND EMOTION CATEGORIES [12]

Category	AU	Category	AU
Satisfied	12,25	Sadly, disgusted	4,10
Unhappy	4,15	Fearfully angry	4,20,25
Frightful	1,4,20,25	Fearfully surprised	1,2,5,20,25
Annoyed	4,7,24	Fearfully disgusted	1,4,10,20,25
Amazed	1,2,25,26	Angrily disgusted	4,25,26
Disgusted	9,10,17	Disgusted surprised	1,2,5,10
Happily sad	4,6,12,25	Happily fearful	1,2,12,25,26
Happily surprised	1,2,12,25	Angrily disgusted	4,10,17
Happily disgusted	10,12,25	Awed	1,2,5,25
Sadly fearful	1,4,15,25	Appalled	4,9, 10
Sadly angry	4,7,15	Hatred	4,7,10
Sadly surprised	1,4,25,26	-	-

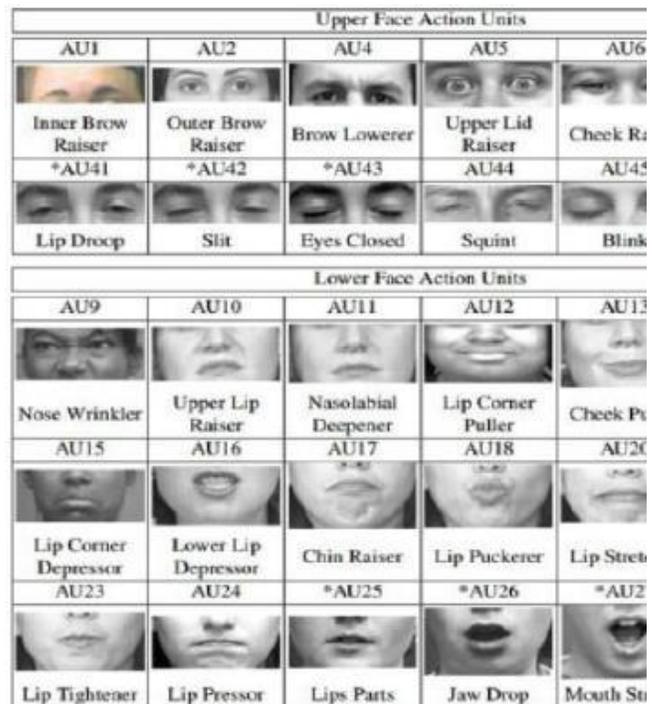


Fig. 1. Different Units of Action for the Upper and Lower Part of the Face [11].

In their paper [14,15], Cha and Kim proposed the use of webcams to measure the duration of attention and a learner's movements. The changes of these facial features were analyzed

in [14]: face up, face down, face turned, eyes closed, eyes open. It is claimed that they "determine the focus and unfocus states for the face and the eye from the coordinates of the extracted characteristic points". The authors fail to explain or define how they achieved this "focused state" or "unfocused state". When the learner's face is in a "front face" position, they conclude they are concentrated. Additionally, they claim that whether a student is "focused" or "non-focused" depends on the data obtained from "the learner's face tilts or turns sideways or their eyes open or close". However, they do not demonstrate a degree of concentration or explain how this analysis is performed. Adding a set of new features to [14], [15], the authors added smiles, surprise, sadness, anger, and closed and open mouths. In the study, "concentration" was defined as "the state of an open eye, open mouth, depressed expression, face turned, and facial expressions of emotion," while "non-concentration" was defined as "the state of a closed eye, open mouth, or dejected expression".

If both the criterion value and the value are above 0.9 milliseconds ", the learner's eyes are closed, indicating that he is not focused. The blinking eye occurs if the value is under 0.9 ms", indicating concentration. Those who do not concentrate will have their eyes closed, their mouth open, their faces turned, or facial expressions of emotion such as smile, surprise, sadness, or anger.

Students' gaze tracking behavior in front of a computer screen is studied by Yi et al. [16]. The resulting learning status is calculated. By assessing this learning status, the quality of the teaching can be evaluated. During the cognitive learning process, the eyes perform three basic actions: scanning, searching, and seclusion. According to them, pupils do not use the information captured by their eyes in a situation of inactivity on a cognitive level. This method, however, only utilizes eye movement information.

Yale and JAFFE [17], Xia Mao and Zheng Li designed an intelligent online learning system to learn about a student's emotional states through facial expression, speech, and text.

Lan Li, Li Cheng, and Kun-xi Qian explored the use of affective computing in facial expression-based e-learning system, and classify learners' emotional states into four categories such as surprise, confusion, frustration and confidence [19].

An analysis and representation of facial dynamics is described in [20]. Using facial expressions, the algorithm calculates the optical flow to determine the direction of movement.

By automatically detecting subtle changes in expressions, [21] is aimed at developing an optical flow-based approach to capturing facial expressions.

Fig. 2 illustrates the comparison of Sanchez et al.'s [22] two FER methods, which are based, respectively, on feature point tracking and dense flow tracking.

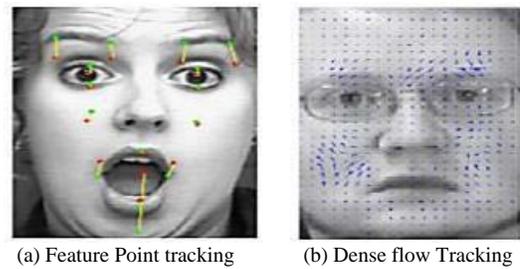


Fig. 2. Applications of Optical Flow based Methods on Facial Images in [22].

When using deep learning for FER, CNN is well suited to detect DU. FACS-based CNN-based FER methods have shown the capacity to generalize both cross-task and cross-data networks associated with FER [23]. Microexpressions are detected by the model in a well-executed manner. By using the CNN of Kim et al. [24], the LSTM is trained to learn the temporal characteristics of a spatial representation through facial expressions. We determine the most representative expressions in facial sequences regardless of the intensity or duration of their expressions as part of network learning. Kritika [25] monitors the position of pupils' eyes and heads, and generates an alert if concentration is low. Videos were analysed and dissected. MATLAB was used to implement and detect faces and Violas-Jones features using different functions. Students can find out whether they are feeling negatively in e-learning environments by using the system. Even though considerable progress has been made in this field of research, emotion and focus still need further exploration. This article attempts to establish an index of student concentration by analyzing facial expressions.

III. PROPOSED METHOD

In this article, a system is proposed that automatically discovers the learner's state of concentration in real time from facial expressions using a web camera.

From the learner's camera, facial expressions are analyzed to determine the student's state of concentration, but there is no standard database for the teaching environment and most studies are based on databases of posed expressions based on six universal expressions such as happiness, sadness, anger, surprise, neutral and fear. These are not suitable for an online learning environment.

The datasets used in this research were collected from different datasets. Therefore, The Concentration State Ranking System consists of three modules, as shown in Fig. 3:

A. Face Detector

This algorithm detects and extracts the student's features quickly and efficiently by using the Haar Cascade algorithm developed by Viola & Jones. In recent years, this method has become one of the most popular methods to accomplish this purpose [26].

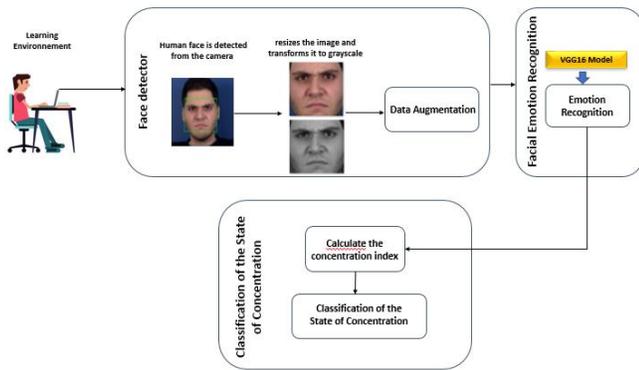


Fig. 3. System Proposed for Determining the Concentration Level of Students in Real Time from their Facial Expressions.

B. Recognition of Facial Emotions

An alternative model (that is useful for detecting facial expressions) uses transfer learning to identify the dominant emotion represented by a student's face at any given moment. There are six categories of facial expressions based on the emotion expressed: angry, scared, happy, sad, surprised, or neutral. Transfer learning is a machine learning approach that focuses on the ability to apply relevant knowledge from previous learning experiences to a different but related problem. We used a transfer learning approach to create a phrase recognition framework using databases (CK+, fer2013, and JAFEE). We used 4 pre-trained models (VGG16, VGG19, Resnet50, AlexNet, Xception) [27], which are deep convolutional networks designed for object recognition [28] and have shown good results on the ImageNet dataset [29] for object recognition. We replaced the last fully connected layers of our models with a dense layer with six outputs. The number of outputs in the last dense layer corresponds to the number of classes to be recognized. We trained the last dense layer with images from the database using the softmax activation function and the ADAM optimizer [30].

C. Classification of Concentration

The concentration index (CI) is calculated by multiplying the probability of dominant emotions (DEP) by the corresponding emotion weights (EW) see Table II. Another way to put it is that emotional weight is the measure of how well one's mental state reflects one's concentration at a given time. A value between 0 and 1 is assigned to it.

According to the results, a student's concentration level is classified into one of three categories: highly concentrated, nominally concentrated, and not concentrated.

TABLE II. CLASSIFICATION OF CONCENTRATION

Emotion	EW
neutral	0.9
happy	0.6
surprised	0.5
sad	0.3
fear	0.3
angry	0.25

- Very concentrated: a student falls into this category when the value of his concentration index based on facial emotion is between 50% and 100%.
- Nominally concentrated: a student falls into this category when the value of the Facial emotion concentration index is between 50% and 20%.
- Not at all concentrated: the student's concentration is in this category when the value of the facial emotion concentration index is less than 20%.

Fig. 3 shows an example of a real-time system that provides information data to teachers in real-time. Teachers and e-learning systems use these data to monitor learners in real time as they stream content, so that the teacher can adjust the teaching accordingly, that is, if the concentration level gets too low score, then the teaching material is too difficult for the learners and the difficulty level of the material can be adjusted.

IV. COLLECTING AND PREPROCESSING DATA

To ensure that the output was not biased in favor of a particular dataset, multiple datasets were gathered.

The following standard facial databases are available online: CK & CK+ [31], FER2013 [32], and JAFEE [33]. There were 29,207 images in the training dataset prior to the increase in data.

- Class 'anger' contains 4160 images,
- Class 'fear' contains 4204 images,
- Class 'happy' contains 7453 images,
- Class 'neutral' contains 4995 images,
- Class 'sadness' contains 4945 images, and the class 'surprise' contains 3450 images.

The validation dataset contained a total of 3,533 images.

- Class 'anger' contains 467 images,
- Class 'fear' contains 496 images,
- Class 'happy' contains 895 images,
- Class 'neutral' contains 607 images,
- Class 'sadness' contains 653 images, and the class 'surprise' contains 415 images.

Examples of model datasets are shown in the following Fig. 5.

A. Face Detection and Cropping

Detecting the location of faces from images is known as the face detection process, or face registration. The faces from the images were detected using OpenCV Cascade [34]. The face was detected and then cropped to avoid the complexity of the background, thereby improving the efficiency of the model.

B. Converting to Grayscale

Red, green, and blue channels were added to the images to make them 224 x 224 pixels. By converting the images to

Grayscale with only one channel, we were able to reduce the pixel complexity in the dataset [18] [35]. The training process was streamlined as a result.

C. Image Augmentation

The amount of data could be increased to improve the model's performance. An image augmentation process produces additional images by performing certain operations on existing image data sets, such as random rotation, zoom, shear, flip, etc. (see Fig. 4).



Fig. 4. Dataset Samples.



Fig. 5. Data Augmentation.

V. EXPERIMENTAL RESULTS

A. Comparison between the Models

Using Keras and TensorFlow (<https://keras.io/>) [15], we created a neural network and image processing system using a Python-based neural network API. The program offers many functions and models, which are important for improving the quality of images.

It provides easy-to-use tools for creating custom neural networks, which facilitate rapid experimentation. Google Colaboratory, a free cloud service that supports GPUs, was used to reduce training time.

To train our dataset, we used 100 epochs of 871 steps each. We analyzed the pre-trained models, VGG-16 and VGG-19, Xception, Alexnet.

The VGG 19 model performed excellently with a training accuracy of 90% (show Fig. 8 and Fig. 9) achieved in the 70th epoch, while the VGG16 and Xception (show Fig. 10, Fig 11 and Fig. 12 and Fig 13) model achieved its full 90% accuracy in the 50th epoch. However, the AlexNet (show Fig. 6 and Fig. 7) model was a bit long with a 91% accuracy in the 90th epoch. The accuracy of Alexnet training was 99.8%, but it could have been higher.

From Table III, we can observe that Vgg16 obtains the highest accuracy 100%. Vgg16 and Vgg19 get the most optimal scores in terms of Accuracy and error rate - 100%, 19% and 99.8%, 18%, respectively.

As a conclusion, VGG16 is a powerful deep learning model for facial emotion classification, specifically for CK+, JAFEE, and FER13 images classification. VGG16 produced the highest training and testing accuracy compared to other models. Vgg19, outperforming AlexNet, obtained Xception the second highest accuracy.

B. Determine the Level of Concentration of Students

The concentration level can be detected by analyzing the result of the seven emotions. The level of concentration can be categorized into three levels: high, medium and low, respectively.

TABLE III. COMPARISON BETWEEN THE MODELS

Model name	Maximum training accuracy	Maximum Validation accuracy
Xception	100%	85%
Vgg16	100%	86%
Vgg19	99.8%	87%
AlexNet	99.8%	64.5%

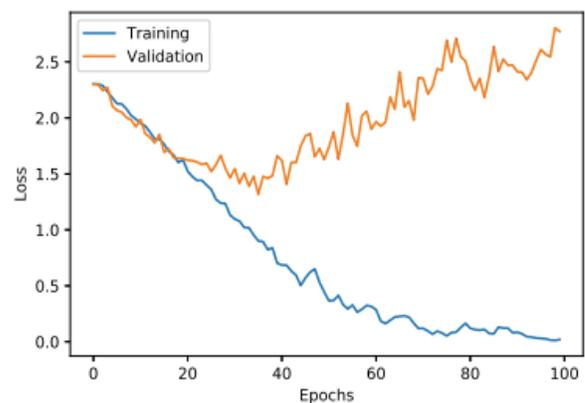


Fig. 6. AlexNet Training and Validation Accuracy Loss.

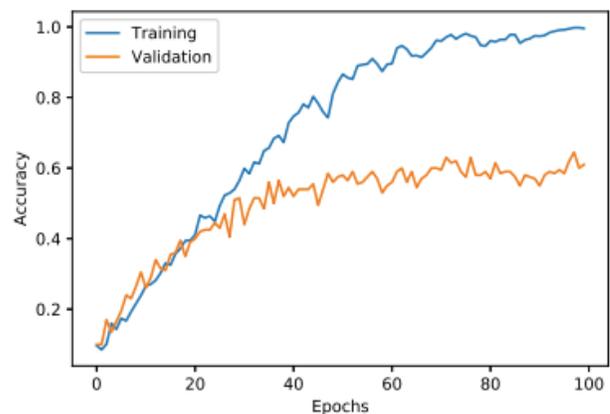


Fig. 7. AlexNet Training and Validation Accuracy.

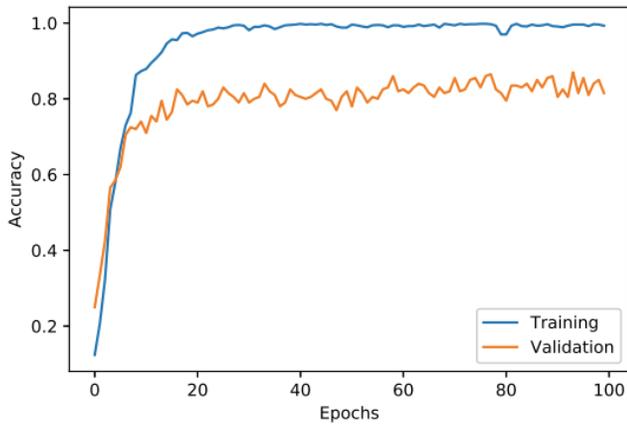


Fig. 8. VGG19 Training and Validation Accuracy.

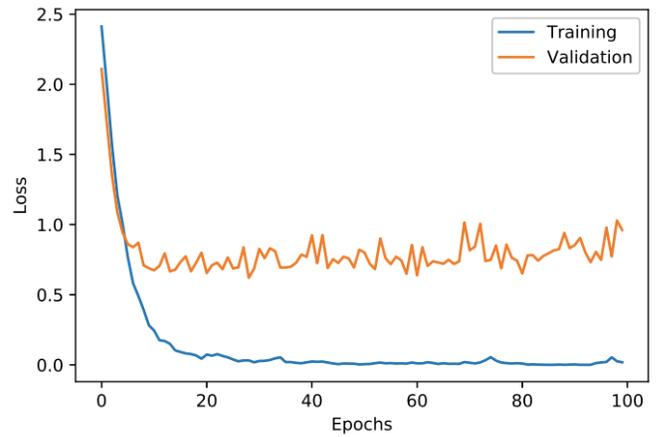


Fig. 11. VGG16 Training and Validation Loss.

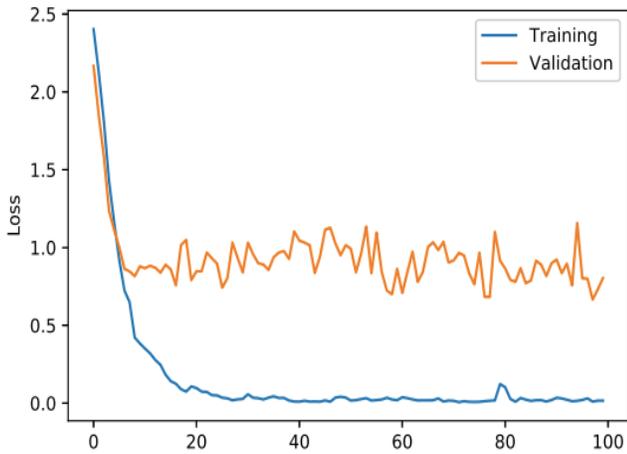


Fig. 9. VGG19 Training and Validation Loss.

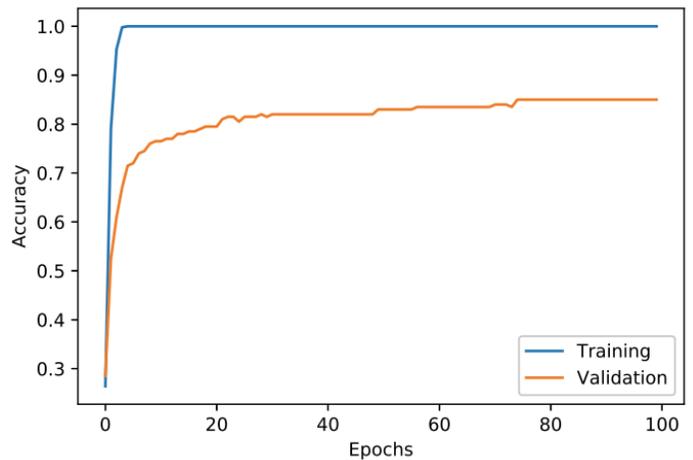


Fig. 12. Xception Training and Validation Loss.

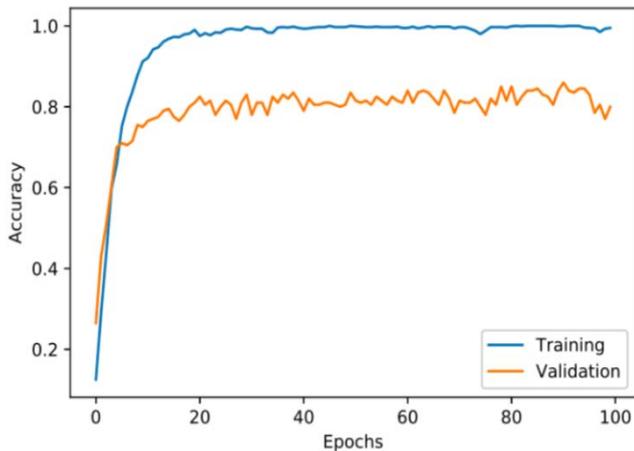


Fig. 10. VGG16 Training and Validation ACCURACY.

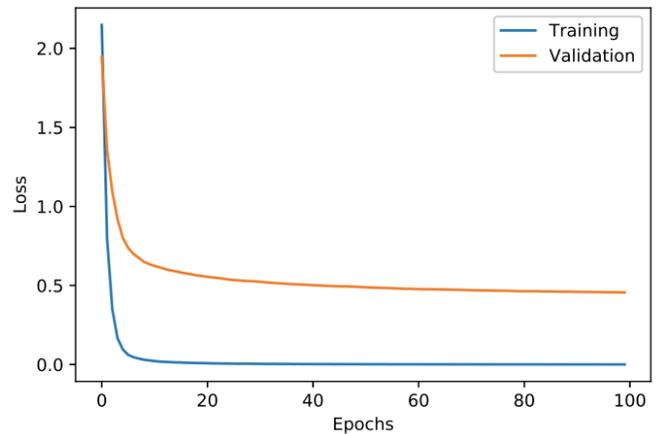


Fig. 13. Xception Training and Validation Accuracy.

An evolution of the student's concentration based only on facial expressions. For the index of concentration to be calculated, each emotion adds its own value. Based only on facial emotion detection, we calculated the percentage concentration using the following rules shown in Table IV:

Fig. 14 shows an overview of the system operating in real time and presenting information data to the teacher. This data, which includes the learner's facial emotions, and the level of concentration detected (Very concentrated, Nominally concentrated, Not at all concentrated).

TABLE IV. CALCULATE THE LEVEL OF CONCENTRATION

Emotion	CI
neutral	$(DEP * 0.9) * 100$
happy	$(DEP * 0.6) * 100$
suprised	$(DEP * 0.5) * 100$
sad	$(DEP * 0.3) * 100$
fear	$(DEP * 0.3) * 100$
angry	$(DEP * 0.25) * 100$
Not Emotion	0 (i.e., Not at all concentrated)

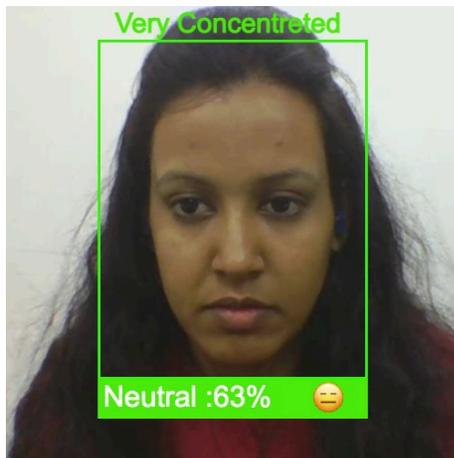


Fig. 14. General view of the System Operating in Real Time.

VI. CONCLUSION

Recognizing human emotions is one of the most challenging tasks in e-learning. In this paper, we explored emotion detection to predict students' concentration level. One of the biggest challenges and benefits of distance learning, and especially online learning, is having a system that can determine the concentration levels of students. We experimented with transfer learning models on the dataset we combined (JAFEE + FER2013+CK+) for facial expression recognition to determine the concentration level of students during educational tasks and then estimated their effectiveness. The results show that VGG16 performs better than other proposed models, while the results show that the VGG19 model could also achieve decent accuracy in facial expression recognition. In our research we present an approach to a system for detecting students' concentration level from facial expressions. Only the web camera provides information to the system. We developed a system that produces a concentration index based on the facial expressions captured by the camera. It was designed to work in real time. Three different concentration levels are presented: "highly concentrated", "nominally concentrated", and "no concentration whatsoever". This system can also help teachers to know the learning state of their students, so that the teacher can adapt the teaching material accordingly, i.e. if the concentration level is too low, the teaching material is too difficult for the learners and the difficulty level of the material can be adjusted. Furthermore, the research can be improved by including an analysis of the history of emotions detected over a given period of time in

order to establish a predictive model of when students are likely to drop out or fail in a subject.

REFERENCES

- [1] Jacob Whitehill, Zewelanj Serpell, Yi-Ching Lin, Aysha Foster, and Javier R Movellan. The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, 5(1):86–98, 2014.
- [2] Michael K Barbour and Thomas C Reeves. The reality of virtual schools: A review of the literature. *Computers & Education*, 52(2):402–416, 2009.
- [3] Neil A Bradbury. Attention span during lectures: 8 seconds, 10 minutes, or more?, 2016.
- [4] Reed W Larson and Maryse H Richards. Boredom in the middle school years: Blaming schools versus blaming students. *American journal of education*, 99(4):418–443, 1991.
- [5] Paul Ekman, Wallace V Friesen, and Phoebe Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*, volume 11. Elsevier, 2013.
- [6] Margaret M Bradley, Laura Miccoli, Miguel A Escrig, and Peter J Lang. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, 45(4):602–607, 2008.
- [7] Hayes, D.: ICT and learning: lessons from Australian classrooms. *Comput. Educ.*2(49),385–395 (2007).
- [8] Arkorful, V., Abaidoo, N.: The role of e-learning, the advantages and disadvantages of its adoption in higher education. *Int. J. Educ. Res.*2(12), 397–410 (2014).
- [9] Yewale, P., Zure, S., Awat, A., Kale, R.: Emotion recognition using image processing. *Imperial J. Interdisciplinary Res.*3(5) (2017).
- [10] Ekman, P.: Universal facial expressions of emotions. *California Mental Health Res. Digest*8(4), 151–158 (1970).
- [11] Du, S., Tao, Y., Martinez, A.: Compound facial expressions of emotion. *Proc. Natl. Acad. Sci. U.S.A.*111(15), 1454–1462 (2014).
- [12] Benitez, Q.C., Srinivasan, R., Martinez, A.: EmotioNet: an accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 5562–5570 (2016).
- [13] Jonathan Bidwell and Henry Fuchs. Classroom analytics: Measuring student engagement with automated gaze tracking. *Behav Res Methods*, 49:113, 2011.
- [14] Cha, S., Kim, W.: Analyze the learner's concentration using detection of facial featurepoints. *Adv. Sci. Technol. Lett.*92,72–76 (2015).
- [15] Cha, S., Kim, W.: The analysis of learner's concentration by facial expression changes & movements. *Int. J. Appl. Eng. Res.*11(23), 11344–11349 (2016).
- [16] Yi, J., Sheng, B., Shen, R., Lin, W., Wu, E.: Real time learning evaluation based on gazetracking. In: 14th International Conference on Computer-Aided Design and Computer Graphics, Shanghai, pp. 157–164 (2015).
- [17] Sathik, M.M. and Sofia, G. Identification of student comprehension using forehead wrinkles. *IEEE International Conference in Computer, Communication and Electrical Technology (ICCCET)*, 2011, 66–70.
- [18] Mao, X. and Li, Z. Implementing emotion-based user-aware e-learning. *ACM in CHI'09 Extended Abstracts on Human Factors in Computing Systems*, 2009, 3787–3792.
- [19] Li, L., Cheng, L. and Qian, K.X. An e-learning system model based on affective computing. *IEEE In International Conference on Cyber worlds*, 2008, 45–50.
- [20] Yacoub, Y.; Davis, L.S. Recognizing human facial expressions from long image sequences using optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.*1996,18, 636–642.
- [21] Cohn, J.F.; Zlochower, A.J.; Lien, J.J.; Kanade, T. Feature-point tracking by optical flow discriminates subtle differences in facial expression. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 14–16 April 1998; p. 396.

- [22] Sánchez, A.; Ruiz, J.V.; Moreno, A.B.; Montemayor, A.S.; Hernández, J.; Pantrigo, J.J. Differential optical flow applied to automatic facial expression recognition. *Neurocomputing* 2011, 74, 1272–1282.
- [23] Breuer, R.; Kimmel, R. A deep learning perspective on the origin of facial expressions. *arXiv* 2017, arXiv:1705.01842.
- [24] Kim, D.H.; Baddar, W.; Jang, J.; Ro, Y.M. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. *IEEE Trans. Affect. Comput.* 2017.
- [25] LB Krithika. Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. *Procedia Computer Science*, 85:767–776, 2016.
- [26] Paul Viola and Michael Jones. 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*, volume 1, pages I–I. IEEE, 2001.
- [27] K. Simonyan, A. Zisserman, Very deep convolutional networks for largescale image recognition, in: *Arxiv*, 2014.
- [28] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (6) (2018) 1452–1464.
- [29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *Computer Vision and Pattern Recognition*, 2009.
- [30] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, *CoRR* arXiv:1412.6980.
- [31] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,” in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.
- [32] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis, “Deep learning approaches for facial emotion recognition: A case study on fer-2013,” in *Advances in Hybridization of Intelligent Methods*. Springer, 2018, pp. 1–16.
- [33] Lyons, M.; Akamatsu, S.; Kamachi, M.; Gyoba, J. Coding facial expressions with gabor wavelets. In *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 14–16 April 1998; pp. 200–205.
- [34] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” in *IEEE*, 2001, p. 511.
- [35] M. Grundland and N. A. Dodgson, “Decolorize: Fast, contrast enhancing, color to grayscale conversion,” *Pattern Recognition*, vol. 40, no. 11, pp. 2891–2896, 2007.

Medical Image Cryptanalysis using Adaptive, Lightweight Neural Network based Algorithm for IoT based Secured Cloud Storage

M V Narayana, Ch Subba Lakshmi, Rishi Sayal
Guru Nanak Institutions Technical Campus
Hyderabad, India

Abstract—Currently available modern medical system generates large amounts of data, such as computerized patient data and digital medical pictures, which must be kept securely for future reference. Existing storage technologies are not capable of storing large amounts of data efficiently. It is a key and abrogating topic of specialized, social, and medical significance and a key and abrogating subject of general interest. The results of cars, purchasers, and Internet of Things industry-based and essential segments, sensors, and other daily objects are fused with a network of the Internet and solid information abilities that promise to alter the way we operate and live future. The suggested work demonstrates a symmetric-key lightweight technique for secure data transmission of images and text, which uses an image encryption system and a reversible data hiding system to demonstrate the program's implementation. On the other hand, cloud storage services can meet demand due to features such as flexibility and availability. Cloud computing is enabled by amazing internet innovation as well as cutting-edge electrical equipment. Even though medical images may be stored on the cloud, most cloud service providers only save client data in plain text. As part of their overall strategy, cloud users must take responsibility for protecting medical data. Because attackers' increasing computing power and creativity are opening up more and more areas in this mathematical form, most existing image encryption schemes are vulnerable to the plaintext attack of choice. This article presents an image encryption method inspired by an Adaptive IoT-based Hopfield Neural Network (AIHNN) that can resist other assaults while optimizing and improving the system through continuous learning and updating.

Keywords—Cloud storage; IoT; medical image; neural network

I. INTRODUCTION

There have been enormous improvements in Information Technology (IT) and analytics during the last several years. Cloud computing is one kind of technology that many businesses use for the hosting of their apps and data storage. Modern gadgets, which allow for digitalized data exchange, have established a new standard for transmitting information globally, regardless of the application. The information is shared with the public. There is a concern about the level of privacy and security that should be provided for people and confidential material [14]. Information sharing across different companies, particularly medical imaging system users, has increased in recent years, resulting in novel storage methods such as cloud edge [1]. Many online services, such as e-

commerce, telemedicine, electronic payment access, and social networking sites are hosted on the cloud.

Storage as a Service (SAS) is by far the most frequently demanded service according to most Internet of Things (IoT) devices [2], and computing has reached a pinnacle in recent years. Data security in cloud storage is a joint responsibility of customers and service providers. Thus, issues related to medical data privacy, in particular, must be successfully handled via the use of suitable cryptographic standards. It is essential for patient safety to be able to offer information security for a medical image stored in a public and shared system such as the cloud. According to the cloud data security debate, cloud service providers are focusing on security measures for cloud infrastructure, hosts, and data. User data, on the other hand, is only accessible as plain text up to a certain point in time.

To address this issue, it is suggested that data be encrypted and stored utilizing encryption methods. The authors proposed a fully homomorphic encryption (FHE) technique to encrypt photos in cloud storage in order to guarantee image security, even for small-scale images, while the complexity increased even for small-scale images [13]. There are many encryption methods available, and they may be used to encode the text, images, and audio. Despite this, the security measure used for text transmission less performs when used for image transmission due to the intrinsic characteristics of images, including such mass data capacity, also has and strong association among the pixel data; as a result, a separate security system is presented by each type of multimedia data protection. Image security becomes more critical due to the difficulties in managing pictures as contrasted to text documents. As a bonus, pictures are used to convey the context of the majority of critical applications such as telemedicine and education and biometric verification and identification. Chaos-based picture encryption methods [9] have gained popularity due to their increased vital strength, which is achieved via key sensitivity. That perhaps the key has an actual value is the reason for the importance of the turmoil. According to chaos theory, the nature of chaos is entirely dependent on the original seeds.

According to cryptanalysts, chaos-based cryptographic protocols, on the other hand, are not resistant to the chosen plaintext attack. On the other hand, most contemporary cryptographic methods are susceptible to cryptanalysis, in which the attackers undermine the cryptosystem by using

known-plaintext and chosen-plaintext assaulting strategies. The reverse exclusive-OR procedure has been integrated into picture encryption because of its advantages, including reversibility and the ability to cause bitwise confusion. It is also possible to construct a stream cipher using this method. It does, however, allow for cryptanalysis via the use of a specific plaintext attack. In this paper, the homomorphic encryption method is discussed as superior encryption technology, emphasizing cloud data security. On the other hand, the homomorphic encryption method is susceptible to cryptanalysis, wherein security keys may be recovered with less than 8s [19].

According to the initial study, most current schemes are vulnerable to plaintext attacks because of their daily use of simple XOR-based propagation and continuous operation rounds. The number of activities may enhance the complexity, but the change in processing time is unneeded, resulting in a low throughput rate. The encryption method must be complicated to avoid cryptanalysis, particularly the selected plaintext attack, and that it is irreversible, self-adaptive, and parameter aware. As previously stated, neural-based encryption methods are the preferred option for meeting the criteria. An artificial neural network (ANN) is a set of predictive classification networks capable of performing multiple classifications and optimization operations in parallel while still including essential components known as neurons. Because of the self-learning and adaptable character of ANN, it may be used in conjunction with traditional methods to provide correct results. Furthermore, ANN may learn about its environment by using real-time and training data. As a consequence, neural networks are being used in a variety of applications, including data protection, predictive analytics, medical data classification, and civil structure analysis [6][7], with data security being the focus of this study.

A. Artificial Neural Networks and Cloud Computing

Artificial neurons should have been able to replicate the activity of actual neurons, according to the basic principle. As a result, it is characterized by erratic behavior. Because of the security concerns, cryptography applications [12] are attracted to the chaotic activity of neurons. According to the research, the ANN model may be expanded to describe complex reversible encoders as well. ANN may also be used in combination with a non-traditional image encryption technique. The ANN may be thought of as a nonlinear encoder that might replace the traditional diffusion method. Random indexes, in addition to dispersion, are needed to achieve the appropriate degree of confusion. To produce random indices, the neural network must also show recurrent activity, which is inevitable in the generation of pseudo-random sequences for picture encryption applications.

As a result, the recurring as a major component of this paper's proposed neural blended adaptive image encryption system, an adaptive IoT-based Hopfield neural network (ANN) is presented (NBAIE). The Adaptive IoT-based Hopfield neural network is an influence exerted based on human brain features [3]. It is a one-of-a-kind neural network model that is based on human mind features. It has a time-dependent behavior. In many respects, it varies from previous neural architectures. Other neural networks are ideally suited for

classification and grouping tasks and other activities since they are made up of distinct hidden units that process the inputs.

In contrast, recurrent AIHNN contains hidden units that are linked. Time-dependent behavior is achieved by activating one of the hidden units at a certain point in time. It is helpful in applications dependent on the sequence of consecutive occurrences, like pseudo-random sequence generation, where the sequence of succeeding occurrences is essential. The following are the key differentiators of the algorithm suggested:

Because of its BPN's multi-layered architectural design and nonlinear activation device weight matrix, predicting the key is reduced.

Because distinguishing features of the image are used as input for the BPN, and the generated keys are much more highly adaptable to the input plain image than the primary image itself.

- Keys' behavior may be varied due to BPN's ability to self-learn.
- Because of its recurrent and chaotic behavior, AIHNN necessitates the use of crucial seeds which are similar to chaos, which is favourable to linear neural architectures including such BAM and BPN because the preliminary essential seed is larger in scale (size is much larger than basic image), reducing communication rate [20].
- Establishing a link between the authorized user and the public cloud computing environment Each image has its own key. As a consequence, unauthorized people will not be able to attack the image and key using a specific plaintext attack.
- Image specific pseudo-sequence creation, followed by dynamic ambiguity and diffusion, to achieve the desired effect.
- Medical picture repositories on the cloud may now have enhanced privacy protection.
- The weight matrix of the AIHNN may be modified for each picture, resulting in the creation of image-specific pseudo sequences using the algorithm.

As a result, the algorithm's prediction becomes more complicated due to the use of an Adaptive IoT-based Hopfield neural network, which is implemented in this scheme [16].

Further, this work is furnished such as in Section – II, describes about Internet of Things are discussed, in Section – III, the parallel research outcomes are analyzed and discussed for finding the bottlenecks of the current applications, in the Section – IV, the identified drawbacks and the proposed solutions are presented using The Proposed Adaptive Iot-Based Hopfield Neural Network (AIHNN), based on the mathematical models, the proposed algorithms are furnished and discussed, the obtained results from these two novel algorithms are furnished and realized, in the Section – V, and in the Section – VI, the final research conclusion is presented.

II. THE INTERNET OF THINGS (IoT)

The Internet of Things (IoT) is expected to provide social and financial benefits to emerging and developing countries soon. Incorporating Blockchain technology addresses supported agriculture, water quality utilization, human services, and ventures managing condition, among other things [5]. The Internet of Things (IoT) also promises to become a means of achieving the Sustainable Development Goals of the United Nations. It is not new for developed countries to be confronted with the enormous scope of Internet of Things problems [10]. The benefits of the Internet of Things must also be considered by the municipalities that are developing them.

More requirements and challenges in putting this concept into action in less-developed areas must be addressed, such as frameworks, marketplace. Enterprise motivating factors, specialist skills, and approach resources, among other things, because the connections between things, the environment, and the general public are becoming more dynamic [4]. The Internet of Things today ensures that we will live in a progressive, fully connected, dazzling world]. However, the tests and concerns regarding the Internet of Things should be examined. Moreover, it should be addressed to determine the possible benefits to individuals, communities, and businesses [17]. Finally, increasing the benefits of the Internet of Things while simultaneously reducing the security risks cannot be achieved by engaging in an unending debate that pits the affirmation of the Internet of Things against its flaws. It will pledge dedication and collaboration throughout the partner meetings to provide the most acceptable ways to the public in the future. The use of Internet of Things segments prompted a slew of legal questions and gave rise to previously unaddressed legal problems relating to the Internet of Things. The inquiries are massive in scope, and the rapid evolution of IoT technology concerns the ability of the associated approach, legal framework, and regulatory zones to be changed. In this group of problems, one would be the information stream which occurs when IoT devices collect information about people within their purview and transmit it to a different location with different data security rules to advance the treatment of the information. Another issue is that the data collected by IoT devices are susceptible to misuse, which results in unsatisfactory outcomes for the customers [11]. Other legal problems associated with the Internet of Things devices include disagreements between law enforcement agencies and standards-setting organizations, information pulverization, and legal obligations for non-required uses, security, and protection concerns. The Internet of Things is now accessible because we have information internet associations, which allows us to access it. The development of information and internet association may be traced back through history and across time. The time information is sent to the Internet facilitates the efficient delivery of information to the Internet. After that, the following section provides a brief historical overview of the information era [18-24].

This study aims to develop and evaluate lightweight and asymmetric block cipher-based cryptosystems that are suitable with MANET, IoT, and wireless communication devices to achieve data security while also maintaining quality control. The following are some possible discussion points on the

work's goal. Prepare a design and simulation of a data security system that incorporates a block cipher, image processing, and reversible data concealing. To develop standard protocols that is interoperable with WSN, MANET, and the Internet of Things. Understanding and implementing the mathematical modeling of cryptographic algorithms is essential for success.

III. PARALLEL WORKS

When it comes to security, a lightweight authentication model for the Internet of Things (IoT) provides a high degree of protection against various assaults such as impersonation attacks, man-in-the-middle attacks, and unknown key sharing attacks in the E-health domain. Based on IoT-based E-health apps, the author proposed a safe, lightweight authentication method that is easy to implement. The suggested approach, based on the Internet of Things, offers authentication, an energy-efficient system, and computing for healthcare. Elliptic curve cryptography (ECC) is a concept that defines the characteristics of the suggested model, in addition to the individuals who offer healthcare and the patients. In general, the author's motivation is to design a lightweight security scheme based on ECC principles for E-health applications based on the IoT (IoT) [8]. The author devised an authentication method based on the minor key, which offers a high degree of security for the system. They also developed a high-performance, lightweight security scheme for E-health applications based on the Internet of Things. The proposed security model is based on the RSA cryptographic algorithm. The algorithm for public-key cryptography is most widely used. In communication stacks, it is used to offer UDP/IPv6 networking to use less power and energy [8].

In [15], a protocol for robust and secure authentication in the Internet of Things systems was suggested to be efficient and safe. In order to ensure security, the proposed protocol uses a physical function that cannot be copied. The proposed protocol protects various attacks while being highly efficient in memory, computations, energy, and communication. An Internet of Things (IoT) mutual authentication mechanism was described by the author. The system is based on PUFs, which use a challenge-response mechanism to send authentication information [15]. Because the protocol offers secure authentication and creates a session key without the need for it, an IoT device does not need to store anything. The author demonstrated that the proposed protocol is highly efficient and provides security against a variety of attacks, including physical attacks, side-channel attacks, and cloning attacks, among others. One of the most important requirements for Internet-of-Things (IoT) systems is security while using as few resources as feasible. Because IoT devices are low-cost and simple, they are a great target for physical, side-channel, and cloning threats, among others. To solve the same issue, the developer developed an effective protocol for mutual authentication for Internet of Things devices.

IV. PROPOSED ADAPTIVE IoT-BASED HOPFIELD NEURAL NETWORK (AIHNN)

AIHNN features metastable states that are triggered by external input and a previously selected state. The chaotic behavior is provided by an ANN that is constructed with a minimal number of nodes, chosen node connections, and a

suitable asymmetric weighted route. The proposed work will be utilized for the AIHNN, which consists of eight nodes. It features an AIHNN architecture with eight nodes, and each node may be connected to any other node and any external input sources. It also has a link to itself. The starting state of the input nodes and external input, as well as other variables, influence the output of each node. The weighted route links every node in the network to every other node. As demonstrated in the previous section, Fig. 1 depicts a recurrent AIHNN that has been rebuilt into a chaotic architecture with eight nodes.

Each node in chaotic design is considered as an input/output node, resulting in a faster generation of the pseudo-random sequence than in traditional architecture. As a result, it is said to be an instance of hyperchaotic architecture. Because each node in this design is not connected to every other node and weights are modified using the Hebb rule in combination with the hyperbolic activation function, this architecture has several distinguishing features. Furthermore, AIHNN has been designed with a certain number of nodes and an appropriate asymmetric weighted path to produce the desired chaotic behavior in a controlled setting.

$$W_{ij} = \begin{bmatrix} w_{11} & w_{12} & w_{13} & w_{14} & w_{15} & w_{16} & w_{17} & w_{18} \\ w_{21} & w_{22} & w_{23} & w_{24} & w_{25} & w_{26} & w_{27} & w_{28} \\ w_{31} & w_{32} & w_{33} & w_{34} & w_{35} & w_{36} & w_{37} & w_{38} \\ w_{41} & w_{42} & w_{43} & w_{44} & w_{45} & w_{46} & w_{47} & w_{48} \\ w_{51} & w_{52} & w_{53} & w_{54} & w_{55} & w_{56} & w_{57} & w_{58} \\ w_{61} & w_{62} & w_{63} & w_{64} & w_{65} & w_{66} & w_{67} & w_{68} \\ w_{71} & w_{72} & w_{73} & w_{74} & w_{75} & w_{76} & w_{77} & w_{78} \\ w_{81} & w_{82} & w_{83} & w_{84} & w_{85} & w_{86} & w_{87} & w_{88} \end{bmatrix} \quad (1)$$

$$w_{ij} = \begin{cases} \sum_{m=1}^M d_i^m d_j^m & \text{if } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

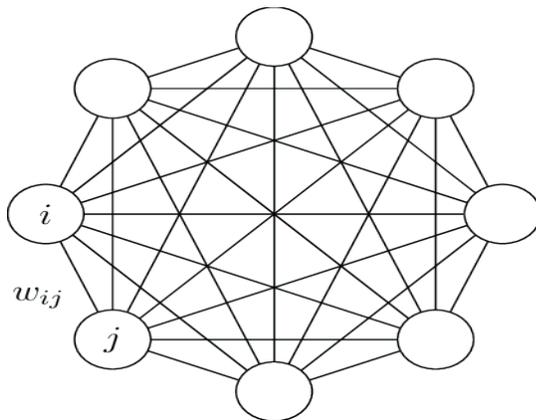


Fig. 1. Architecture of Hyperchaotic AIHNN with Eight Nodes.

Given that the selected Adaptive IoT-based Hopfield neural network has eight nodes, the weight matrix dimension is 8x8, equivalent to W11–W88 in the preceding example. The weight

values in Equation 2 relate to the neural network strength acquired during training, which influences the accuracy of the produced result (rate of closeness among required and received output). The weight values are either integers or floating decimal numbers, depending on the activation function (identity or hyperbolic activation function). The following equation (2) and the following equation (3) are used to incrementally update each node (3). Each node in the preceding equation (2) receives a weighted signal from the other nodes, as well as external input and information from other nodes (Xi). The adjusted sigmoid transfer function is then computed using the revised Xi as shown in the equation:

- The architecture depicted in Fig. 1 produces cyclic random sequences by deriving the following equations: (2) and (3), where Cmax is the steady and Max.
- Cmax is the initial state.
- Wij is the mass function in both.

$$M_{\max} = \frac{n}{2 \log(n)} \quad (3)$$

$$C_{\max} = \frac{n^2}{2 \log(n)} \quad (4)$$

$$Q = \frac{n(n-1)}{2} \log_2(P) \quad (5)$$

Adaptive IoT based Hopfield Neural Network (AIHNN)

$$P_{\max} = M_{\max} + 1 \text{ and then}$$

$$Q_{\max} = \frac{n(n-1)}{2} \log_2(M_{\max} + 1) \quad (6)$$

This leads to:

$$\eta = \frac{n}{(n-1) \log(n) \log_2 \left(\frac{n}{\log(n)} + 1 \right)} n \quad (7)$$

$$\approx \frac{1}{\log(n) \log_2 \left(\frac{n}{\log(n)} \right)} n \quad (8)$$

V. EXPERIMENTAL RESULTS AND DISCUSSION

Visual information that is not medical is referred to as Image data (I), and it serves as a carrier of Critical Medical Information (CMI). As shown in Fig. 2, a Secret Key (SK) is utilized by the Steganographic embedding function (SFE) to conceal CMI, and Stego data (SD) is produced as an output (by the device at the transmitting end DT).

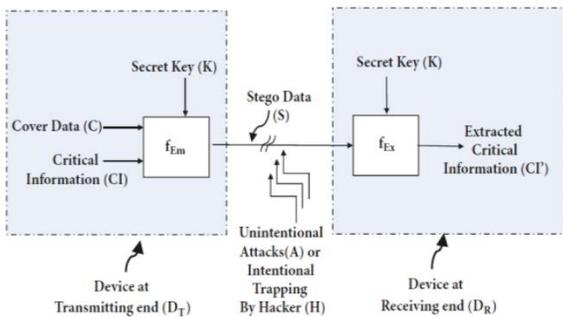


Fig. 2. Steps Involved in Reversible Hiding Algorithm.

The Steganographic extraction function (SFE) extracts CI using the precise Secret Key (K) provided by the user (as a device at receiving end DR). The stego data is CI is Extracted Critical Information, and the secrete key is the secrete key. Fig. 3 depicts a generalized hardware Steganographic data concealing device that is used in the industry. Crypto xStegoSystem is the name given to the proposed reversible data concealing system, intended for use in the implementation of a data hiding system and comprises cryptographic and Steganographic methods. The suggested approach is shown in the Fig. 3.

The central idea of the investigation may be broken down into the many shown in Fig. 3.

- 1) Ensure that low-complexity symmetrical key encryption is implemented on various platforms (JPEG-2000, JPEG, BMP, PNG, GIF).
- 2) Evaluation of the performance of the simulated method on user-defined and real-time imagery.
- 3) An investigation of incorrect key encryption.
- 4) Design and development of the Graphical User Interface for the proposed system.
- 5) An evaluation of the suggested system's execution time and memory allocation is performed.
- 6) The development of a reversible data-hiding system for picture and text multiple encryptions is underway.
- 7) Evaluation of the suggested system's AIHNN and MSE results as shown in Fig. 4 and Table I.

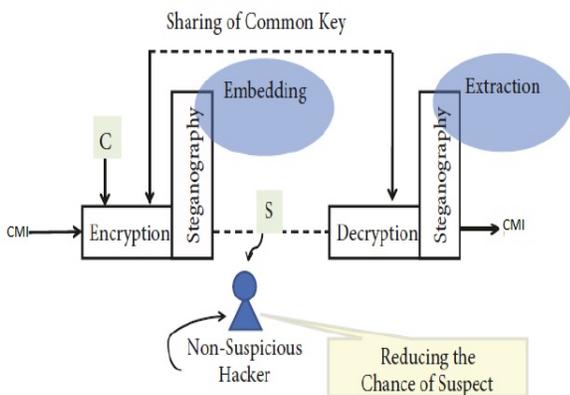


Fig. 3. Phases in Key Sharing in AIHNN Algorithm.

Algorithm 1: The overall process is explained as follows: -

1. Validate CI and store it in the Message Cache
2. Accept the LOOK-UP Table as an Embedding Key in Step 2. (K)
3. Parse the cover data in 8-byte chunks (C)
4. Compute the DWT of the first 8bytes of the cover data
5. Select a byte(Ym) at random from the Message Cache.
6. It was chosen using 3LSBs of the contents of the selected byte LOOK-UP Table.
7. Insert the chosen bit at DWT coefficient C3 to complete the operation.
8. Compute IDWT of 8bytes of cover data in order to get Stego information.
9. Repeat steps three within 8 for all of the bits in the Message Cache and all of the characters.

TABLE I. AIHNN ENCRYPTION AND DECRYPTION MSE

Type of Medical Image	ANN	MSE
BMP	70.543	0.0242
GIF	66.231	0.0242
JPEG	67.001	0.0231
JPEG-2000	71.643	0.0176
PNG	68.332	0.0209

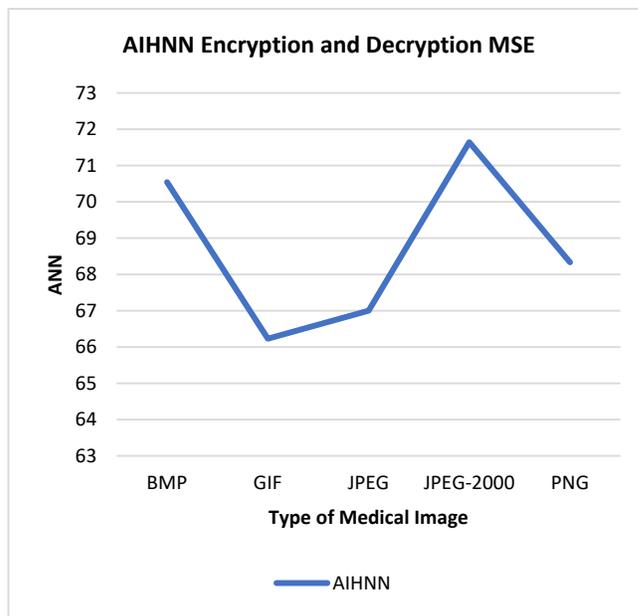


Fig. 4. AIHNN Encryption and Decryption MSE.

The method presented here is to create chaotic sequences using AIHNN rather than nonlinear equations that show chaotic behavior. The image-specific key is treated as a control parameter, and the bit of color is produced using the pseudo-code shown in Algorithm 2. Followed by the tested images and its testing results mentioned in Fig. 5, Fig. 6, Fig. 7 and Table II.

Algorithm -2: Random sequence generation using AIHNN and image specific key

Input: Multiplicative identity matrix (B) $_{(8 \times 8)}$, Sampling rate T_w , Random initiator $h_{(1 \times 8)}$

Output: Nonlinear random sequence Ω

Step 1. Initialize

$$w_{ij} \leftarrow \left[\frac{w}{2\sigma} - w; \Psi 2w 3w 0; 3w \phi w 0; MW w 0 0 nw \right] \\ \leftarrow [1 \quad 0.5 \quad -5 \quad -1 \quad ; -0.37$$

Step 2. $2 \quad 3 \quad 0 \quad ; 3 \quad -13 \quad 1 \quad 0 \quad ; 100 \quad 0 \quad 0$

Update the w_{ij} with new σ, Ψ, ϕ

Get $H(0) \leftarrow [h]^T$

Step 3. For i, j 11 to do

$$f(H(r)) = \tanh H(r)$$

$$H(r + 1) = (1 - BT_w)H(r) + T_w wf(H(r))$$

$$Dh = |(H(r + 1) - [|H(r + 1)|])|$$

Step 4. Until $r \leq (\text{Image size} / 4)$ Initialize $\Omega \leftarrow \{ \}$ for $f \leftarrow 1$ to 16384

Step 5. for $i \leftarrow 1$ to $4 \Omega((4 \times (f = 1)) + i) = Dh(i)$

Step 6. End

Step 7. End

Step 8. Return Ω

The encrypted medical image has also been stored in public cloud storage, and only authorized customers will be forced to access the clouds to get a ciphered image. The main image may only be accessed using the private key(s) that were assigned to it.

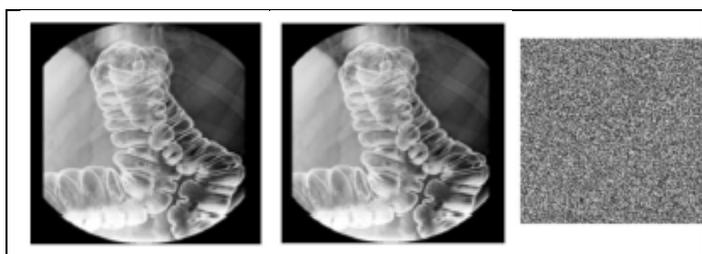


Fig. 5. Medical: (i) A Plain Image (Img); (ii) an Encrypted Image (EImg = E(Img)); (iii) A Decrypted Picture (Img = D(EImg)).

In order to arrive at exclusive encrypted pictures for each medical image, the proposed AIHNN system changes the weight matrix for each medical image. The adaptive encoding aspect of the proposed work has resulted in an average entropy of 8.11, independent of the original medical pictures used in the analysis.

Healthcare terminology:

Keys are important to the functioning of encryption systems, and key sensitivity analysis is a helpful tool for evaluating the robustness of encryption methods. Analysis of key sensitivities. For the most part, encryption techniques

utilise the same key for each image transmission. The proposed approach, on the other hand, offers an independent and adjustable key for each image based on image attributes. When the key is in double data type, it is feasible to reflect even minor changes in the image's properties.

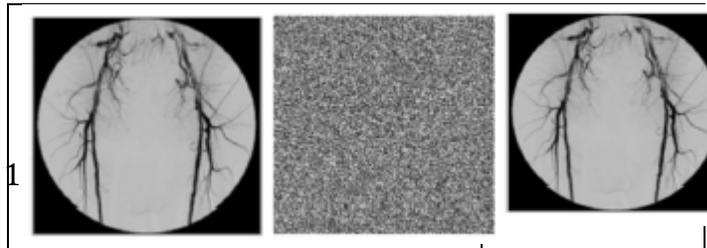


Fig. 6. (iv) Plain (MImg); (v) Encrypted (EImg = E(MImg)); (vi) Decrypted (MImg = D(EImg)).

TABLE II. TEST IMAGES WITH ENTROPY

Test images	Global Entropy		Local entropy
	Original image	Encrypted image	No. of blocks =50 Block size: 88*88 Encrypted image
MI1	4.5	8.1245	8.1106
MI2	5	8.1263	8.1112
MI3	6.3	8.1272	8.1118
MI4	6.5	8.1287	8.1124
MI5	7.4	8.13005	8.1131
MI6	8.13	8.1314	8.1136
MI7	8.86	8.13275	8.1142
MI8	9.59	8.1341	8.1148
MI9	10.32	8.13545	8.1154
MI10	11.00	8.1368	8.1163

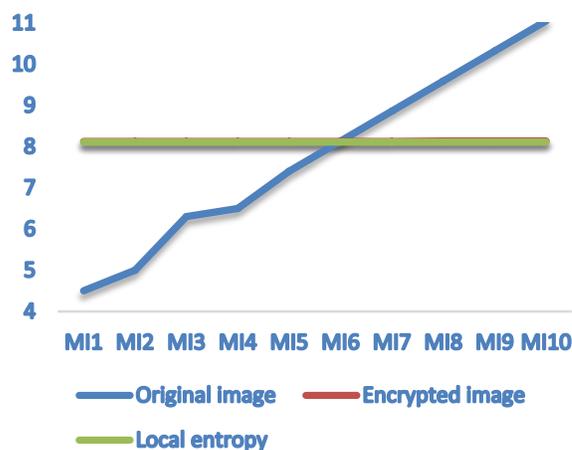


Fig. 7. Test Images with Entropy.

VI. CONCLUSION

The consolidation of computers, sensors, and systems to pass the report on and control devices have been around for a long time, but as significant technological and business trends advance, the "Internet of Things" is experiencing a new

phenomenon known as the "Internet of Things." A progressive, fully linked world is promised by the Internet of Things, connecting different things and their surroundings and between articles and people becoming increasingly securely connected over time. On a basic level, it has the potential to alter how people perceive what it means to be "on the web." Even the potential is significant, and there are many obstacles to overcome – primarily in the areas of security and protection, interoperability and scales, legal problems, and human rights issues, including emerging economies. Consequently, it is necessary to recognize and address its problems while maximizing its benefits and minimizing its risks. Answers for enhancing the optimal use of IoT while also minimizing the risks cannot be found by being involved in a heated debate that pits the benefits of IoT against the concerns of cybersecurity.

There are numerous obstacles to overcome when it comes to the Internet of Things, including intensity, data transmission capacity, flexibility, security, and protection. Predefined security arrangements at each tier are nevertheless vulnerable to security-related attacks, even though they have been strengthened. Because of their openness and multi-tenancy, cloud storage systems are susceptible to a wide range of security vulnerabilities. This suggested solution provides data security for data saved in cloud storage and data in various states such as underuse, rest, and transit, among others. Creating a secure medical picture archive on the cloud is the driving force behind this project. It is crucial to highlight that the Hopfield attractor is an important component of the surveillance system for medical images stored in the cloud, and its fitness is also validated using standard metrics. The proposed research utilized a Hopfield attractor to confuse pixels, followed by diffusion, and the findings demonstrated that the system was resistant to additional attacks.

REFERENCES

- [1] Asadi S, Nilashi M, Husin ARC, Yadegaridehkordi E (2017) Customers perspectives on adoption of cloud computing in the banking sector. *Inf Technol Manag* 18:305–330.
- [2] Dana Halabi, Salam Hamdan, "Enhance the security in smart home applications based on IoT-CoAP protocol.
- [3] Domer B, Fest E, Lalit V, Smith IFC (2003) Combining dynamic relaxation method with artificial neural networks to enhance simulation of tensegrity structures. *J Struct Eng* 129:672–681. [https://doi.org/10.1061/\(ASCE\)0733-9445\(2003\)129:5\(672\)](https://doi.org/10.1061/(ASCE)0733-9445(2003)129:5(672)).
- [4] Himanshu Gupta, GarimaVarshney, "A Security Framework for IoT devices against wireless threats, second international conference on telecommunication and networks, 2017.
- [5] Jin HyeongJeon, Ki-Hyung Kim, "Blockchain-based data security-enhanced IOT server platform, IEEE ICOIN, 2018.
- [6] Kohonen T (1988) An introduction to neural computing. *Neural Netw* 1:3–16. [https://doi.org/10.1016/0893-6080\(88\)90020-2](https://doi.org/10.1016/0893-6080(88)90020-2).
- [7] Maria Almulhim, Noor Zaman, "Proposing secure and the lightweight authentication scheme for IoT based E-health applications" International conference on advance communication technology; 2018.
- [8] Mourad Talbi and Med Salim Bouhleb, "Application of a Lightweight Encryption Algorithm to a Quantized Speech Image for Secure IoT", Preprints.org; 2018. DOI: 10.20944/preprints201802.0096.v1.
- [9] Muhammad NaveedAman, KeeChaing Chua, "A lightweight mutual authentication protocol for IOT system, 2017.
- [10] MuhammetZekeriyaGunduz, Resul Das, "A comparison of cyber security-oriented testbeds for IOT based smart grids, IEEE 2016.
- [11] Nithya C, Pethururaj C, Thenmozhi K, Amirtharajan R (2020) An advanced framework for highly secure and cloud-based storage of colour images. *IET Image Process*. <https://doi.org/10.1049/iet-ipr.2018.5654>.
- [12] Qin K (2017) On chaotic neural network design: a new framework. *Neural Process Lett* 45:243–261. <https://doi.org/10.1007/s11063-016-9525-y>.
- [13] Qin Z, Weng J, Cui Y, Ren K (2018) Privacy-preserving image processing in the cloud. *IEEE Cloud Comput* 5:48–57. <https://doi.org/10.1109/MCC.2018.111121403>.
- [14] Shahzadi S, Iqbal M, Dagiuklas T, Qayyum ZU (2017) Multiaccess edge computing: open issues, challenges and future perspectives. *J Cloud Comput*. <https://doi.org/10.1186/s13677-017-0097-9>.
- [15] Singh A, Chatterjee K (2017) Cloud security issues and challenges: a survey. *J Netw Comput Appl* 79:88–115. <https://doi.org/10.1016/j.jnca.2016.11.027>.
- [16] Tang H, Li H, Yan R (2010) Memory dynamics in attractor networks with saliency weights. *Neural Comput* 22:1899–1926. <https://doi.org/10.1162/neco.2010.07-09-1050>.
- [17] Tao M, Ota K, Dong M (2017) Ontology-based data semantic management and application in IoT- and cloud-enabled Smart homes. *Future Gener Comput Syst* 76:528–539. <https://doi.org/10.1016/j.future.2016.11.012>.
- [18] Dr. Ranga Swamy Sirisati, M Vishnu Vardhana Rao, S Dilli Babu, Dr. M.V.Narayana, "An Energy efficient PSO based Cloud Scheduling Strategy", Lecture Notes in Networks and Systems book series (LNNS, volume 171) Springer, Singapore, Print ISBN 978-981-33-4542-3, Online ISBN 978-981-33-4543-0, DOI: https://doi.org/10.1007/978-981-33-4543-0_79, - pp: 749-760.
- [19] Thomas Maurin, Laurent, George Caraiman, "IoT security assessment through the interfaces P-SCAN test bench platform, 2018 EDAA.
- [20] Yu W, Cao J (2006) Cryptography based on delayed chaotic neural networks. *Phys Lett Sect A Gen At Solid State Phys* 356:333–338. <https://doi.org/10.1016/j.physleta.2006.03.069>.
- [21] U Shivanna, NiladriShekar Dey, K Purnachand, M V Narayana, Govardhana Rao I, "A Comparative Study of Famous Image Compression Methods Based on Bits per Pixel: A Survey", *Journal Of Critical Reviews*, VOL 7, ISSUE 18, 2020, pp.1094-1104, ISSN-2394-5125.
- [22] M V Narayana, "Compression, Encryption, Watermarking & Steganography (CEWS) Technique for Image Steganography" *International Journal of Latest Engineering and Management Research (IJLEMR)*, Volume 3, Issue 3, PP. 20-27, (March, 2018), e-ISSN: 2455-4847– UGC Indexed Journal- 48163.
- [23] M V Narayana, U Shivanna "A Review on Region of Interest (ROI) based compression Techniques for Medical Images" *International Journal of Management, Technology And Engineering*, Volume 7, Issue IV, APRIL/2017 PP 51-56. ISSN NO : 2249-7455 (UGC Approved Journal).
- [24] M V Narayana, U Shivanna "Local Features Based Image Matching Using Sift Algorithm" *International Journal of Research*, Volume 5, Issue 2, 2016 PP 51-55. ISSN NO : 2236-6124 (UGC Approved Journal).

Analysis of Logistics Service Quality and Customer Satisfaction during COVID-19 Pandemic in Saudi Arabia

Amjaad Bahamdain¹, Zahyah H. Alharbi², Muna M. Alhammad³, Tahani Alqurashi⁴
Management Information Systems Department, King Saud University, Riyadh, Saudi Arabia^{1,2,3}
Common First Year Deanship, Computer Science Department, Umm Al-Qura University, Makkah, Saudi Arabia⁴

Abstract—Logistics companies' success is inextricably linked to the quality of their services, particularly when dealing with customer issues. Nowadays, social media is the first place that users turn to in order to express their thoughts on services or to communicate with customer service representatives to resolve problems. Businesses can retrieve and analyze these data to gain a better understanding of the factors that affect their operations, both positively and negatively. During the COVID-19 pandemic, we conducted a sentiment analysis to assess customer satisfaction with logistics services in Saudi Arabia's private and public sectors. Using a lexicon-based approach, 67,124 tweets were collected and classified as positive, negative, or neutral. A support vector machine (SVM) model was used for classification, with an average accuracy of 82%. Following that, we conducted a thematic analysis of negative opinions in order to identify the factors that influenced the effectiveness and quality of logistics services. The findings reveal five negative themes: delay, customer service issues, damaged shipments, delivery issues, and hidden prices. Finally, we make suggestions to improve the efficiency and quality of logistics services.

Keywords—Logistics services; sentiment analysis; lexicon-based approach; SVM; sentiment classification

I. INTRODUCTION

Logistics adds value by meeting customers' delivery needs in a cost-effective manner [1]. As a result, logistics service performance represents a provider's ability to consistently deliver requested products within the specified time frame and at an acceptable cost [2, p. 34]. This means that the quality of the provided logistics services can be considered the most significant success guarantor of a logistics business [3]. It increases customer satisfaction, which in turn increases the business's profitability and competitive advantage [4]. Customer satisfaction will determine whether or not customers make further purchases or refer the business to others [5].

People rely heavily on social media for information, news, and entertainment. Twitter and other social media platforms are growing in popularity because they provide easy access to real-time posts on a daily basis. Tweets and mentions on Twitter contain information about current events, news, user opinions, and reviews. These data are useful to businesses because they allow them to better understand user preferences, improve their services, and increase customer satisfaction. Twitter provides an Application Programming Interface (API) through which researchers, developers, and business owners

can obtain data to analyze and learn more about topics of interest to them. Sentiment Analysis is the process of extracting, analyzing, and distinguishing opinions or emotions from text [6]. It is a branch of natural language processing (NLP) research. Sentiment analysis is currently used to analyze data posted on social media platforms or websites in order to determine opinions, attitudes, or emotions about businesses, products, or services. These classifications are beneficial to business owners because they allow them to identify their strengths and weaknesses in order to improve future services and thus increase profits. To the best of our knowledge, no previous studies have used Twitter users' opinions to analyze customer satisfaction with logistics services in Saudi Arabia during the COVID-19 pandemic. Therefore, this study aims to fill this gap and to achieve the following objectives: 1) to investigate customer satisfaction with logistics services during COVID-19 in Saudi Arabia by analyzing customer opinions; and 2) to identify the elements of logistics service quality that influence customers' sentiments toward logistics service providers.

The following is how the paper is structured. First, we review the relevant literature, and describe the methods employed to collect the data and conduct the analysis. The results are presented, followed by a discussion of the results. Finally, the managerial implications of the findings and future research directions are explored.

II. LITERATURE REVIEW

This section begins by reviewing logistics service quality and customer satisfaction. Second, we discuss the various tools, methods, and results of sentiment analysis studies that have used Arabic tweets. We also review how researchers have used sentiment analysis to investigate customer satisfaction.

A. Logistics Service Quality and Customer Satisfaction

Customer satisfaction is defined as the perceived difference between prior expectations and actual product or service performance [7]. It is a sign that customers' needs are being met in a pleasurable manner. According to some researchers, the ultimate goal of any organization is to meet the needs and achieve the satisfaction of its customers. Understanding the needs and expectations of customers is a necessary step in ensuring their satisfaction [8].

The research of Parasuraman, Zeithaml, and Berry [9] identified five broad dimensions of service quality that influence customer satisfaction: dependability, receptiveness, assurance, empathy, and tangibles. Dependability explains the ability to provide the promised service consistently and reliably. Receptiveness represents the willingness to assist customers and provide prompt service. Assurance symbolizes employee knowledge and courtesy, as well as the ability to convey trust and confidence, while empathy is about providing caring and individualized attention to customers. Finally, tangibles represent the appearance of physical facilities, equipment, personnel, and communications materials. Several studies have applied these dimensions to identify the quality gap in logistics and to improve the services in order to achieve customer satisfaction. For example, [10] conducted a study to validate five determinants of service quality and to investigate the service quality-customer satisfaction link in the port logistics service industry in Vietnam. According to the findings, five factors influence the quality of port logistics services: responsiveness, assurance, dependability, tangibles, and empathy. The findings additionally show that the quality of port logistics services has a positive impact on customer satisfaction.

Furthermore, Cho et al. [11] presented three factors of service quality in order to determine their effects on customer satisfaction, referral intentions, and loyalty. The study was linked to the port logistics service industry in Incheon and Shanghai. The result of their study showed that exogenous and relational quality remained important dimensions of customer satisfaction. According to their findings, relational quality included the positive attitude and professionalism of employees, partnership relations between the customers and the ports, and the ability of the port information system to provide timely information. However, the endogenous quality of services has an insignificant impact on customer satisfaction.

Uvet [12] conducted a study that investigated the aspects of logistics services that affected customer satisfaction. These included timeliness, quality of personal contact, operational information sharing, order condition and handling of order discrepancies. He applied structural equation modeling (SEM) and confirmatory factor analysis (CFA) to examine customer satisfaction by utilizing the five factors of logistics service quality. The results showed that there are significant relationships between logistics service quality factors and customer satisfaction. Additionally, the finding of this study can be applied in any business to obtain competitiveness in logistics services.

B. Sentiment Analysis and Customer Satisfaction

Anastasia and Budi [13] examined Twitter opinion data to estimate users' satisfaction with the GO-JEK and Grab companies, which provide online transportation services in Indonesia. From February to March 2016, they retrieved 126,405 tweets using the Twitter API and the R programming language. RapidMiner was also used for data pre-processing and classification. The study examined the highest accuracy score using three classification algorithms: SVM, Naïve Bayes

(NB), and Decision Tree. The results showed that Grab had higher user satisfaction than GO-JEK, and that the SVM and Decision Tree algorithms had the highest accuracy with a score of 72.97%.

From September 2013 to February 2014, Gitto and Mancuso [14] conducted an exploratory sentiment analysis on passenger review data extracted from a blog called "Airport Reviews" on the SKYTRAX website. Their study looked at customer satisfaction for both aviation and non-aviation airport services. The study looked at five international airports in Europe: Amsterdam Schiphol, Frankfurt, London Heathrow, Madrid Barja, and Paris Charles de Gaulle. The researchers used KNIME, an open-source platform for data analytics, and Semantria, which uses a dictionary base and a machine learning technique, to conduct sentiment analysis and to detect positive, neutral, or negative sentiments. According to the findings, non-aviation services such as food and beverage and the shopping area had the greatest influence on passenger satisfaction levels. Non-aviation services received approximately 55% positive feedback, while aviation services received only 33%.

Bhattacharjya et al. [15] investigated the effectiveness of customer services on Twitter for logistics services provided by e-retailers. The study looked at conversations between customers and businesses rather than individual tweets to investigate the effectiveness of customer services. A large sample of 203,349 tweets from e-retailers was collected, as well as a random sample of logistics conversations. After data preparation and cleansing, a total of 16,998 tweets were analyzed. According to the findings, e-retailers and logistics service providers mostly redirect customers to other channels to solve their problems, while ignoring their customers' needs to find solutions and get their issues resolved on Twitter. The findings also revealed a lack of interaction between e-retailers and logistics service providers over providing effective customer service.

Ahmadi et al. [16] performed a sentiment analysis on public service conversations in a hospital in Yogyakarta, Indonesia. The study created the "Kata Kita" product to translate conversations into text and to help deaf people communicate with service personnel by displaying the text on the screen. Customer satisfaction was also evaluated. The study combined NLP with K-nearest neighbors (KNN) and the term frequency-inverse document frequency (TF-IDF) algorithm to classify conversations as "satisfied" or "dissatisfied". The results showed 74.00% accuracy, 76.00% precision, and 73.08% recall.

This literature review demonstrates that researchers all over the world have been focusing on customer opinions in order to better understand customers and the factors that influence their satisfaction. Furthermore, due to the significance of this type of analysis and the scarcity of studies using sentiment analysis to analyze Arabic tweets to understand customers' satisfaction with logistics services, this study will analyze customers' opinions of logistics services during the COVID-19 pandemic.

C. Arabic Sentiment Analysis

Several studies have used Arabic sentiment analysis to gain a better understanding of various phenomena. Aldayel and Azmi [17], for example, proposed a hybrid approach that combined a lexicon-based technique with machine learning techniques to develop a sentiment analysis of Arabic tweets retrieved via Twitter's API that addressed social issues in Saudi Arabia. They used data cleaning, normalization, stop-word removal, an n-gram model to remove repeated letters, and a light stemmer for Arabic text stemming during the pre-processing phase. The training data were labeled using a lexicon-based classifier, and the classification model was built using an SVM classifier. According to the study's findings, the hybrid approach achieved an F-measure of 84% and an accuracy of 84.01%.

Al-Ghaith [18] created the SaudiSentiPlus, a Saudi dialect lexicon that includes a translation of English sentiment lexicons as well as 7,139 terms manually extracted from Saudi tweets. The dataset's polarity classification was carried out manually by three annotators who were Arabic native speakers and Saudis. They divided the dataset into three categories: positive, negative, and neutral. For each classification, the output was 300 labeled tweets. The study conducted an experiment to assess the performance of the SaudiSentiPlus lexicon, using 971,000 tweets from Saudi dialect hashtags as a testing dataset. The accuracy of their results was 81%, which is considered very good.

Al-Horaibi and Khan [19] proposed a sentiment classification system based on a combination of Arabic and English sentiment lexicons. A total of 14,984 tweets were collected from Twitter using the Twitter API and the Python Tweepy2 library. The tweets were pre-processed by the researchers, who removed @usernames, hashtag signs, symbols, numbers, non-Arabic letters, and re-tweets. Following the cleaning process, a total of 3,200 tweets were produced, and these were annotated by two native Arabic annotators. The researchers only used 2,000 tweets because the remaining tweets created uncertainty. There were 634 positive tweets, 675 negative tweets, 691 neutral tweets, and 26,349 tokens in total. A classification semantic approach that used the Arabic Sentiment Lexicon (ArSneL) and the English SentiWordNet lexicon was then employed. The findings indicated that there was a need to improve Arabic stemmer tools by including more roots to support the search process. In addition, the Arabic spelling correction tools needed to be improved, and more words needed to be added.

Al-Hussaini and Al-Dossari [20] proposed a sentiment analysis approach based on a lexicon that included a Saudi dialect lexicon. The lexicon was built using a semiautomatic building technique, with the lexicon being built manually after the data were collected from Twitter. Then, two lists were

manually constructed: one with the most frequently occurring positive sentiment words; and the other with the most frequently occurring negative sentiment words. The lexicon was divided into two columns: words and their polarities. Following classification, there were 762 positive words/phrases with a +1 value and 662 negative words/phrases with a -1 value. The results show that the developed Saudi dialect lexicon helped to improve the sentiment labeling of Arabic tweets.

III. RESEARCH METHODOLOGY

In this section, we will present our Arabic sentiment analysis methodology, which consists of four phases: data collection, data pre-processing, lexicon-based classification, and sentiment analysis. Fig. 1 illustrates the implemented Arabic sentiment analysis methodology, and the following sections explain each stage in more detail:

A. Data Collection Phase

The Python Tweepy2 library was used to extract tweets from the Twitter API. In total, 67,124 tweets were collected that mentioned public and private logistics companies. These included Aramex, DHL, Fetchr, Naqel, SMSA, and Saudi Post. The retrieved tweets were tweeted during the surge of the COVID-19 pandemic, precisely from June to September 2020.

B. Data Pre-processing Phase

The data pre-processing phase consists of three successive stages: text cleaning, tokenization and stemming. In the text cleaning, a number of steps were performed as follows:

- Removing irrelevant tweets that includes ads and were not related to our task.
- Removing non-Arabic letters, including numbers, symbols, the hashtag sign and emojis as they are not part of the analysis.
- Replacing "@username", "www." and "https?:///" with empty strings.
- Removing the punctuation.
- Eliminating character repetition in a word by replacing it with a single character, such as "ويبيبيبي شحنتنتنتي" which is replaced with "وين شحتني".
- Normalizing some characters, including "إ", "!", and "آ" with "ا", the "ة" character with "ه", the "ى" character with "ي", and the character "ؤ" with "و".
- Removing stop words, the standard Arabic stop words, provided by the Natural Language toolkit (NLTK) library, were used.

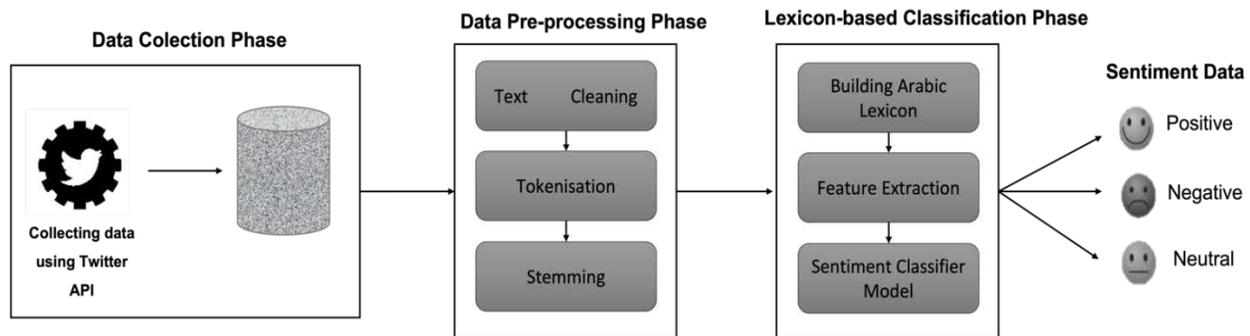


Fig. 1. The Implemented Arabic Sentiment Analysis Methodology.

Then, we split the text into tokens or words using NLTK tokenization¹.

Finally, for the stemming step, we used an Arabic light stemmer called ISRI Arabic Stemmer [21] to remove prefixes and suffixes in order to obtain the root of words. The output of the stemmed words, however, was rooted into three-character words, making the majority of the words unreadable or without meaning. As a result, we did not proceed with the stemming step. According to Albraheem and Al-Khalifa [22], stemming may be ineffective for Arabic dialects as most words lack roots or have roots with more than three characters. Another study [23] confirmed that Arabic dialect does not receive adequate processing in Arabic sentiment analysis because Arabic dialectal varieties necessitate advanced pre-processing and lexicon-building steps.

C. Lexicon-based Classification Phase

For sentiment analysis, a lexicon-based approach was used. First, an Arabic lexicon was needed to label the data before applying the classifier algorithm. There are few Arabic lexicons available, and it is difficult to find one that includes Saudi dialects. Only two studies, [24], [25], were discovered that constructed the Arabic lexicon with Saudi dialects in mind. Therefore, we updated the lexicon created by [18] for categorizing tweets as “positive” or “negative” by including words related to logistics services in each list. The resulted lexicon contained approximately 6,500 words. The lexicon was then applied to assign polarity scores to each word in the tweets: either 1, 0, or -1 according to whether they were positive, neutral, or negative, respectively. Following that, each tweet was labeled based on its maximum polarity scores, and when a tie occurred it was labeled as neutral. For instance, if a tweet contained three negative and two positive words, it was labeled as negative, and vice versa. It was otherwise labeled as neutral.

Second, the feature extraction step was performed on our dataset. This step aims to extract numerical features from text data to make it more suitable for a classification algorithm. We used the TF-IDF algorithm, which increases the weight of rare words, while diminishing the weight of words that occur frequently in the dataset.

After that, SVM was used to build the sentiment classifier model. SVM makes use of a learning algorithm that is based on statistical learning theory and optimization theory. It enables the computer to learn how to perform classification and regression tasks, to improve prediction accuracy, and to avoid the drawbacks of overfitting [26]. The main idea behind SVM is to find the hyper-plane that divides document vectors into two distinct classes [27]. According to several previous studies, SVM performs well with Arabic text. For example, [28] tested three different classification algorithms, the NB, SVM and KNN classifiers, and they found that the SVM classifier outperformed the other classifiers in terms of the precision measure. Another study confirming the ability of SVM to classify Arabic tweets was [29]. The authors tested six different classification algorithms: SVM, KNN, NB, Neural Network, and two Decision Tree algorithms including J48 and C5.0. They concluded that SVM outperformed the other tested classifiers. Moreover, they showed that using the SVM classifier without any stemming on Arabic text was better than other stemming methods [23]. Therefore, it was deemed to be suitable for this research.

D. Experimental Setup

We used the train test split () function from the Python library SciKit-learn to split the dataset into 70% for training and 30% for testing during the training phase. The data were then subjected to a 10-fold cross-validation to ensure their accuracy and to prevent overfitting. The data were divided into ten parts in the cross-validation method. The first run included nine parts for training and one for testing. Another part was used for testing in the next run, while the remaining nine parts were used for training. The final accuracy is the mean of the accuracies obtained from the ten runs. The SVM classification model then predicted the sentiments of the tweets using the predict function to put the classification model to the test.

IV. SVM PERFORMANCE EVALUATION

To assess the performance of the SVM classifier, we employed a variety of evaluation metrics that can fairly judge the model and assess its performance. Accuracy, recall, and precision were calculated and Table I presents these results for SVM while Table II shows classification performance by sentiment polarity. The average accuracy of the SVM classifier in this study was 82%. The results show that SVM has a superior ability in 3-class classification because it can distinguish among classes of sentiment polarity.

¹NLTK (Natural Language toolkit) tokenization is the process of dividing large text into smaller parts called tokens.

TABLE I. RESULT OF SVM MODEL

Classifier	Accuracy	Precision	Recall	F1
SVM	0.82	0.81	0.80	0.80

TABLE II. RESULTS OF CLASSIFICATION PERFORMANCE BY SENTIMENT POLARITY

Classifier	Polarity	Precision	Recall	F1
SVM	Neutral	0.74	0.73	0.73
	Positive	0.82	0.80	0.81
	Negative	0.86	0.88	0.87

V. THEMATIC ANALYSIS RESULTS

After identifying positive and negative tweets, we conducted a thematic analysis of negative opinions in order to identify the factors influencing the quality of logistics services. We used Python to search for codes in the tweets and to then identify the most frequently occurring themes. As a result, five negative opinion theme categories emerged. These themes are: customer support issues, damaged shipments, delay, delivery issues, and hidden prices. Table III displays the negative themes, along with the percentage of occurrences in the entire dataset.

A. Customer Support Issues

Customers reported a variety of problems related to customer support, including a lack of support from customer service representatives and poor service from delivery drivers. For example, 5.7% of the tweets were associated with bad customer service. Many customers complained that customer service representatives were unable or unwilling to resolve their issues – some of the linked tweets are given below (note: translated from Arabic):

“I want to file a complaint about your company, I have been waiting for my shipment for two months and I contacted you and you replied to me, we are working to follow up the request, and I did not see anything.”

“Unfortunately, I submitted a complaint without any concern and care from you.”

“I filed a complaint and the matter was neglected as if it was an unimportant report and they did not contact me till now.”

“I sent complaints till I got bored, yet no one answers.”

TABLE III. NEGATIVE THEMES

Themes	Occurrences	Percentage
Customer Support Issues	2765	5.7%
Delay	1986	4.0%
Damaged Shipments	511	1.0%
Delivery Issues	4736	9.7%
Hidden Prices	756	1.5%

B. Delay

Approximately 4% of the negative tweets showed negative attitudes toward delayed shipment. Customers complained that their shipments were either late or had never arrived. Some examples of these complaints are as follows:

“I hope the problem will be solved, two months have passed, and my shipment has not arrived yet.”

“My shipment was supposed to be shipped on August 1, but it was not shipped, and you were 10 days late for the supposed delivery date.”

C. Damaged Shipments

About 1% of the negative tweets were related to receiving damaged shipments. Customers complained that their shipments had been damaged, broken, opened, or expired. Here are some examples of these tweets:

“I received my shipment very late. And I received it with damaged products, spilled products, and the shipment arrived in a very bad condition ... Who will compensate me for the damage?”

“65 days delay and the shipment arrived opened and dirty.”

“I inform you that my shipment arrived after delaying three days of calling today at 1pm, the temperature was 47 degrees Celsius. The car was not air-conditioned, and the box was at a high temperature, noting that the shipment contains vitamins, shampoo and makeup that does not bear the heat, and it has been spoiled.”

D. Delivery Issues

About 9.7% of the data showed negative opinions about delivery issues. Customers mentioned that they faced many issues with the delivery driver while delivering their shipment. Many customers complained that the delivery drivers were unreachable. Some examples of customer’s complaints are as follow:

“A company like you should employ respectable employees who know how to deal with customer ... unfortunately your delivery driver was very bad and disrespectful today. He calls and raises his voice and talks with impoliteness to deliver the shipment! It is your duty to deliver the shipments to their owners!!”

“Respond to me first and find my shipment!! My shipment has arrived two months ago, and the delivery driver did not answer my calls, and his mobile was locked.”

“I shipped my shipment through express mail and paid extra money to deliver the shipment faster ... in the end I did not receive the shipment for two weeks.”

Customers also complained about missing items in their shipments, lost shipments, delivering to the wrong location, and failing to provide the promised door-to-door service. See the examples below:

“I’ve called and you told me that my shipment was lost, and I’ve contacted you 3 months ago, and I’ve been told that you will communicate with me, but nothing happened.”

“Unfortunately, they charge large fees for shipping shipments, and deliveries are delayed and their shipments arrive incomplete! I have a shipment about two weeks ago, and today it arrives incomplete.”

“A paid shipment arrived two weeks ago and arrived at the wrong address. I contacted you 3 times to change the location and deliver it to my home, and there is no response!”

“The delivery driver left my shipment with negligence at the end of the street and not in front of the door of the house and went, which indicates the lack of integrity of the company representative!”

E. Hidden Prices

Customers were dissatisfied with shipment companies’ hidden fees. Approximately 1.5% of the data showed negative attitudes toward extra charges and taxes. The following are some examples:

“Your representative asked me to pay an additional delivery fee, or he will return my shipment. I refused to pay, and he refused to deliver.”

“Why pay 30 riyals for the delivery fee? I’ve paid for the order shipping fee earlier!”

VI. DISCUSSION AND RECOMMENDATIONS

Based on the above results, it is clear that logistics service providers in Saudi Arabia need to improve their service quality to maintain their credibility and reputation, and to gain customers’ trust. The domain of logistics services in Saudi Arabia is still immature, although significant changes have taken place in the last decade. Logistics service providers should try to gain competitive advantages over others by providing high-quality services that distinguish them from others [12].

Communication between the customer and employees is highly important during service delivery in terms of meeting customers’ expectations [12]. Logistics service providers should offer a variety of customer support mechanisms, such as instant messaging with chatbots to answer frequently asked questions, change the location details, or reschedule delivery dates [30]. It is also critical to provide real-time support with customer service representatives in order to resolve customer issues [31], [32]. In addition, logistics service providers should provide temperature-controlled vehicles and dry gel packs for fragile shipments in order to avoid damage to goods [33].

Logistics service providers should make certain that customers are always kept up to date on the status of their shipments and any delays. Radio Frequency Identification (RFID) can be used by logistics service providers to provide real-time data about shipments, allowing them to know exactly where the shipment is from start to finish. RFID can improve shipment visibility, prevent shipment loss, prevent delivery to the incorrect destination, and reduce customer frustration caused by shipment delays by providing them with

a real-time delivery plan. RFID also reduces the amount of effort and time required to resolve these issues [34].

To resolve issues with unclear fees, transparency should be increased by creating an informed list that shows the shipping process, information required from customers, and a detailed bill. This procedure will save time and money while also increasing customer satisfaction [35]. Additionally, to mitigate the impact of delayed shipments on customer satisfaction during peak times, logistics service providers can offer customers discount coupons as a form of compensation for the delays and to reduce their disappointment and anger [36]. With regard to delivery drivers, logistics service providers’ systems should allow customers to rate their experience with the delivery drivers. This would allow the company to identify and control drivers’ behavior, improve the service quality, and ensure customer satisfaction [37]. In addition, the system could increase employee loyalty by providing bonuses to the highest-ranking employees [38].

VII. CONCLUSION

In this paper, we have evaluated the quality of logistics services during the COVID-19 pandemic in Saudi Arabia by analyzing customers’ opinions. Sentiment analysis was performed to analyze customers’ opinions expressed via the Twitter platform. Specifically, an SVM classifier was applied to predict the sentiment polarity of customers’ opinions; this resulted in 82% accuracy, 81% precision, and 80% recall. Our findings revealed that the majority of the opinions were negative. Further investigation using thematic analysis revealed a number of themes representing factors that had a negative impact on the quality of logistics services. Based on the findings, we developed a set of recommendations to help logistics service providers improve their service quality. We intend to expand the lexicon by adding more Saudi dialects in the future, as well as broadening our approach to analyzing customer reviews on apps. This will allow us to assess companies’ strengths and weaknesses and to assist them in improving the quality of their services via their apps.

REFERENCES

- [1] T. P. Stank, T. J. Goldsby, and S. K. Vickery, “Logistics service performance: Estimating its influence on market share,” *J. Bus. Logist.*, vol. 24, no. 1, pp. 27–55, 2003.
- [2] Bowersox, Donald J., David J. Closs, and M. Bixby Cooper (2002), *Supply Chain Logistics Management*, New York: McGraw-Hill/Irwin, p. 34.
- [3] I. Meidutė-Kavaliauskienė, A. Aranskis, and M. Litvinenko, “Consumer satisfaction with the quality of logistics services,” *Procedia - Soc. Behav. Sci.*, vol. 110, pp. 330–340, 2014, doi: <https://doi.org/10.1016/j.sbspro.2013.12.877>.
- [4] A. Sharma, D. Grewal, and M. Levy, “The customer satisfaction/logistics interface,” *Journal of Business Logistics.*, vol. 16, no. 2, p. 1, 1995.
- [5] J. J. Cronin and S. A. Taylor, “Measuring Service Quality: A Reexamination and Extension,” *J. Mark.*, vol. 56, no. 3, pp. 55–68, Jul. 1992, <https://doi.org/10.1177/002224299205600304>.
- [6] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and trends in information retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [7] F. Buttle, “Customer relationship management: concepts and technology,” *Sydney: a Butterworth-Heinemann*, vol. 1, 2009.
- [8] D. K. Tse, F. M. Nicosia, and P. C. Wilton, “Consumer satisfaction as a process,” *Psychology & Marketing*, vol. 7, no. 3, p. 177, 1990.

- [9] A. Parasuraman, V. A. Zeithaml, and L. L. Berry, "Servqual: A multiple-item scale for measuring consumer perc," *J. Retail.*, vol. 64, no. 1, pp. 12–40, 1988.
- [10] D. N. Le, H. T. Nguyen, and P. H. Truong, "The Asian Journal of Shipping and Logistics Port logistics service quality and customer satisfaction: Empirical evidence from Vietnam," *Asian J. Shipp. Logist.*, vol. 36, no. 2, pp. 89–103, 2020, doi: 10.1016/j.ajsl.2019.10.003.
- [11] C.-H. Cho, B.-I. Kim, and J.-H. Hyun, "A comparative analysis of the ports of Incheon and Shanghai: The cognitive service quality of ports, customer satisfaction, and post-behaviour," *Total Qual. Manag. Bus. Excell.*, vol. 21, no. 9, pp. 919–930, Sep. 2010, doi: 10.1080/14783363.2010.487677.
- [12] H. Uvet, "Importance of logistics service quality in customer satisfaction: An empirical study," *Oper. Supply Chain Manag. An Int. J.*, vol. 13, no. 1, pp. 1–10, 2020.
- [13] S. Anastasia and I. Budi, "Twitter sentiment analysis of online transportation service providers," in 2016 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2016, pp. 359–365, doi: 10.1109/ICACSIS.2016.7872807.
- [14] S. Gitto and P. Mancuso, "Improving airport services using sentiment analysis of the websites," *Tour. Manag. Perspect.*, vol. 22, pp. 132–136, Apr. 2017, doi: 10.1016/j.tmp.2017.03.008.
- [15] J. Bhattacharjya, A. Ellison, and S. Tripathi, "An exploration of logistics-related customer service provision on Twitter: The case of e-retailers," *Int. J. Phys. Distrib. Logist. Manag.*, vol. 46, no. 6–7, pp. 659–680, Jul. 2016, doi: 10.1108/IJPDLM-01-2015-0007.
- [16] S. Ahmadi, S. Shokouhyar, M. H. Shahidzadeh, and I. Elpiniki Papageorgiou, "The bright side of consumers' opinions of improving reverse logistics decisions: A social media analytic framework," *International Journal of Logistics Research and Applications.*, 2020, doi: 10.1080/13675567.2020.1846693.
- [17] H. K. Aldayel and A. M. Azmi, "Arabic tweets sentiment analysis - A hybrid scheme," *J. Inf. Sci.*, vol. 42, no. 6, pp. 782–797, Dec. 2016, doi: 10.1177/0165551515610513.
- [18] W. Al-Ghaith, "Developing lexicon-based algorithms and sentiment lexicon for sentiment analysis of Saudi dialect tweets," *International Journal of Advanced Computer Science and Applications(IJACSA)*, 10(11), 2019. <http://dx.doi.org/10.14569/IJACSA.2019.0101112>.
- [19] L. Al-Horaibi and M. B. Khan, "Sentiment Analysis of Arabic Tweets Using Semantic Resources," *Int. J. Comput. Inf. Sci.*, vol. 12, no. 2, 2016, doi: 10.21700/ijcis.2016.118.
- [20] H. Al-Hussaini and H. Al-Dossari, "A Lexicon-based Approach to Build Service Provider Reputation from Arabic Tweets in Twitter" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(4), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080459>.
- [21] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," *Int. Conf. Inf. Technol. Coding Comput. - Vol. II*, vol. 1, pp. 152–157, 2005, doi: 10.1109/ITCC.2005.90.
- [22] L. Albraheem and H. S. Al-Khalifa, "Exploring the problems of sentiment analysis in informal Arabic," in *ACM International Conference Proceeding Series*, 2012, pp. 415–418, doi: 10.1145/2428736.2428813.
- [23] A. Assiri, A. Emam, and H. Aldossari, "Arabic Sentiment Analysis: A Survey," *International Journal of Advanced Computer Science and Applications(IJACSA)*, 6(12), 2015. <http://dx.doi.org/10.14569/IJACSA.2015.061211>.
- [24] S. Almujaivel, "Covid-19_1M_Saudi_Tweets. GitHub repository," 2020. https://github.com/salmujaivel/Covid-19_1M_Saudi_Tweets (accessed Jun. 04, 2021).
- [25] N. Alrumayyan, S. Bawazeer, R. AlJurayyad, and M. Al-Razgan, "Analyzing user behaviors: A study of tips in foursquare," in *Advances in Intelligent Systems and Computing*, vol. 753, pp. 153–168, 2018. doi: 10.1007/978-3-319-78753-4_12.
- [26] N. Cristianini and B. Scholkopf, "Support vector machines and kernel methods: The new generation of learning machines," *AI Mag.*, vol. 23, no. 3 SE-Articles, p. 31, Sep. 2002, doi: 10.1609/aimag.v23i3.1655.
- [27] G. Fung and O. L. Mangasarian, "Incremental support vector machine classification." In *Proceedings of the 2002 SIAM International Conference on Data Mining*, pp. 247–260. Society for Industrial and Applied Mathematics, 2002.
- [28] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in *Proceedings - 2014 International Conference on Future Internet of Things and Cloud, FiCloud 2014*, Dec. 2014, pp. 579–583, doi: 10.1109/FiCloud.2014.100.
- [29] W. Alabbas, H. M. Al-Khateeb, A. Mansour, G. Epiphaniou, and I. Frommholz, "Classification of colloquial Arabic tweets in real-time to detect high-risk floods," in 2017 International Conference On Social Media, Wearable And Web Analytics, Social Media 2017, Oct. 2017, vol. 2017-June, pp. 1–8, doi: 10.1109/SOCIALMEDIA.2017.8057358.
- [30] V. Marino and L. Lo Presti, "Engagement, satisfaction and customer behavior-based CRM performance," *J. Serv. Theory Pract.*, vol. 28, no. 5, pp. 682–707, Jan. 2018, doi: 10.1108/JSTP-11-2017-0222.
- [31] Y. Park and S. C. Gates, "Towards real-time measurement of customer satisfaction using automatically generated call transcripts," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 1387–1396, doi: 10.1145/1645953.1646128.
- [32] B. J. Goldenberg, *CRM in real time: empowering customer relationships*. Information Today, Inc., 2008, pp.293.
- [33] C. Sykes, "Time- and temperature-controlled transport: Supply chain challenges and solutions," *P& T: a peer-reviewed journal for formulary management*, vol. 43, no. 3, 2018, pp. 154–170.
- [34] S. Garrido Azevedo and H. Carvalho, "Contribution of RFID technology to better management of fashion supply chains," *Int. J. Retail Distrib. Manag.*, vol. 40, no. 2, pp. 128–156, Jan. 2012, doi: 10.1108/09590551211201874.
- [35] Z. Tian, R. Y. Zhong, A. Vatankhah Barenji, Y. T. Wang, Z. Li, and Y. Rong, "A blockchain-based evaluation approach for customer delivery satisfaction in sustainable urban logistics," *Int. J. Prod. Res.*, vol. 59, no. 7, pp. 2229–2249, 2021.
- [36] D.-S. Chang and T.-H. Wang, "Consumer preferences for service recovery options after delivery delay when shopping online," *Soc. Behav. Personal. An Int. J.*, vol. 40, no. 6, pp. 1033–1043, 2012.
- [37] F. Xiang and W. Wu, "Research on emotional labor and management of takeaway delivery staff," in 6th Annual International Conference on Social Science and Contemporary Humanity Development (SSCHD 2020), 2021, pp. 485–489.
- [38] S. Moro, R. F. Ramos, and P. Rita, "What drives job satisfaction in IT companies?," *International Journal of Productivity and Performance Management*, vol. 70, no. 2, pp. 391–407, 2021, <https://doi.org/10.1108/IJPPM-03-2019-0124>.

Investigate the Ensemble Model by Intelligence Analysis to Improve the Accuracy of the Classification Data in the Diagnostic and Treatment Interventions for Prostate Cancer

Abdelrahman Elsharif Karrar
College of Computer Science and Engineering
Taibah University, Medina, Saudi Arabia

Abstract—Class imbalance problem become greatest issue in data mining, imbalanced data appears in daily application, especially in the health care. This research aims at investigating the application of ensemble model by intelligence analysis to improving the classification accuracy of imbalanced data sets on prostate cancer. The primary requirements obtained for this study included the datasets, relevant tools for pre-processing to identify the missing values, models for attribute selection and cross validation, data resembling framework, and intelligent algorithms for base classification. Additionally, the ensemble model and meta-learning algorithms were acquired in preparation for performance evaluation by embedding feature selecting capabilities into the classification model. The experimental results led to the conclusion that the application of ensemble learning algorithm on resampled data sets provides highly accurate classification results on single classifier J48. The study further suggests that gain ratio and ranker techniques are highly effective for attribute selection in the analysis of prostate cancer data. The lowest error rate and optimal performance accuracy in the classification of imbalanced prostate cancer data is achieved using when Adaboost algorithm is combined with single classifier J48.

Keywords—Ensemble model; intelligence analysis; classification of imbalanced data; prostate cancer

I. INTRODUCTION

This Prostate cancer is among the leading causes of death in men worldwide. The prostate is a glandular structure located in the male productive system and its functions is to promote spermatid health and enhance fertility by adding a nutrient-rich alkaline fluid to the semen [1]. Malignant tumors that lead to prostate cancer state to develop when the rate of cell multiplication is higher than cell death. This alters the genetic structure leading to mutations and tumor metastasis on the urothelial lining. Compared to other glands, the prostate has a higher malignancy rate due to the heavy reliance on the androgenic signaling of hormones such as testosterone, abnormal Gli-1 oncogene expression, and Sonic Hedgehog (Shh) expression, which stimulate cellular proliferation and stromal tumor growth. The process in which prostate cancer develops is known as Prostatic Intraepithelial Neoplasia (PIN) While most research studies on the pathogenesis of prostate cancer report inconclusive findings, etiological factors such as

genetic inheritance and family history, vasectomy, environmental carcinogens, low carotenoid intake, and high intake of saturated fats and other unhealthy dietary/lifestyle habits are known to increase the risks significantly.

Prostate cancer is classified as a carcinoma since its malignancy develops primarily from the epithelium lining of the peripheral glandular tissue. The epithelial structure of the prostate gland is composed of three cell types including rare neuroendocrine cells, basal cells, and luminal cells, which are responsible for the expression of androgen receptors, secretion of glycoprotein prostate specific antigen (PSA) and prostatic fluids [1]. Research studies suggest that prostate tumors that initially form from the luminal cells metastasize more rapidly compared to those from the basal cells due to the alteration of epithelial stromal tissues and the damage of glandular structure. The accuracy of clinical interventions such as the classification of diagnostic data from cancer tissues is influenced by a range of factors including the extent of cellular differentiation on histology and cyclic biochemical recurrence risk. Accurate classification of diagnostic data significantly influences the efficacy of treatment intervention through timely detection based on the Tumor, Nodes and Metastasis framework.

Data classification techniques for the diagnostic data are subject to structural imbalances and errors due to factors such as the underlying assumptions of evenly distributed training datasets. The classification approaches are highly vulnerable to bias when implemented on training data sets with severely imbalanced distribution. Insights from imbalanced training data sets may have severe practical implications on the associated decision outcomes. However, the problem of imbalanced data distribution is fairly common in real-world scenarios, especially when target classes lack uniform distribution across multiple class levels [2]. Data set imbalances occur when major classes have more instances and minor classes have relatively fewer instances. The classification of data sets with imbalanced distributions is a major challenge that has not been fully solved even by advanced machine learning algorithms with mathematical model mapping and computational prediction capabilities for identifying embedded data patterns [3].

This paper develops multiple potential approaches based on algorithmic modification, feature selection, ensemble learning, cost-sensitive learning, and sample selection methods to address the challenges of imbalanced distribution in learning data sets.

A. Production Statement

Class imbalance is the most occurring and potentially risky analytics issue, especially in the data mining of unstructured sets from healthcare systems and processes due to the high likelihood of some classes having larger sample sizes compared to others [4]. A significant number of the current data mining techniques are structurally designed to ignore misclassification risks on minor samples while focusing on the classification of major samples. The accuracy of data classification techniques is impeded by factors such as data imbalances coupled with uneven distribution and sample size differences from one class to another. As a result, traditional classification algorithms are highly unreliable and unsuitable due to high risks of bias and inaccuracy. This explains the need to determine and test whether a data classification model based on machine learning ensemble is capable of delivering comparatively higher levels of accuracy [5].

B. Research Questions

This research seeks to answer the following questions;

- 1) Can the implementation of machine learning ensemble model to data classification improve the classification of imbalanced data sets for prostate cancer management?
- 2) Is the application of resampling techniques based on machine learning reliable in optimizing and improving classification accuracy in imbalanced data sets for prostate cancer management?

This research paper is organized in sections including a review of recently published literature on classifiers and prostate cancer for comparisons with related studies in both fields in Section II, a detailed description of the experimental procedure, methodology, imputation process, and the general set up in Section III, and the evaluation of experimental results in Section IV. Finally, Section V of this research paper discusses conclusions based on the experimental results and provides recommendations for future studies.

II. LITERATURE REVIEW

This section provides a conceptual description of data mining with relation to techniques such as ensemble learning and resampling to investigate the implications of data classification accuracy on the management of prostate cancer, including a review of recently published literature on classifiers and prostate cancer for comparisons with related studies in both fields.

A. Prostate Cancer

According to 2021 prevalence statistics by the American Cancer Society, prostate cancer is the second most prevalence type cancer after skin cancer among men in the United States with approximately 248,530 new reported cases and about 34,130 deaths [6]. Data further shows that one in every 8 men develops prostate cancer, especially among adults aged above

65 of African ethnicity. Prostate cancer is ranked as having the second highest death rate from lung cancer in American males. Statistical estimates suggest that in a sample population of 41, one man dies of prostate cancer [6]. In addition to age, other risk factors for prostate cancer in men include family history through genetic inheritance, ethnicity (60% more risk among blacks), and lifestyle factors such as diet, smoking, and level of physical activity [6]. Early detection of prostate cancer is linked to significantly higher chances of survival and longevity. Studies suggest that timely detection of prostate cancer plays a significant role in the effectiveness of treatment interventions hence the need for various interventions to promote the identification and detection of early symptoms.

B. Data Mining Process

Data mining techniques are applied used to extract trends and patterns through the Knowledge Data Discovery process (KDD) [7]. The extraction of patterns among multiple variables depends on data mining techniques, which may be predictive or descriptive. Predictive data mining methods provide a generalized description of the data attributes while predictive data mining uses historical data to make accurate trend forecast [8].

Data Mining software are designed analyze data on the basis of parameters such as sequence analysis (a pattern in which events are interdependent), degree of association (where events defined by the datasets are interconnected), clustering (where data with identical patterns are grouped), and classification (where predefined variables are used to identify new patterns) [9].

C. Data Mining Techniques

The flow diagram shown in Fig. 1 provides a description of various techniques for data mining and retrieval based on regression, classification, clustering, and association [10].

D. Classification Techniques

The classification approach to data mining in healthcare entails predicting and grouping a data set in sample class categories [11]. This provides important insights for the identification of unique disease patterns that associate certain risk factors to a patient population through supervised learning [12]. Binary classification is a technique where the risk factors are classified as either 'high' or 'low' while multiclass technique involves more than two classes for example 'high', 'medium', or 'low risks' [8]. The data is further divided into classes; training and testing datasets, which are used to predict the possible outcomes from a historical event.

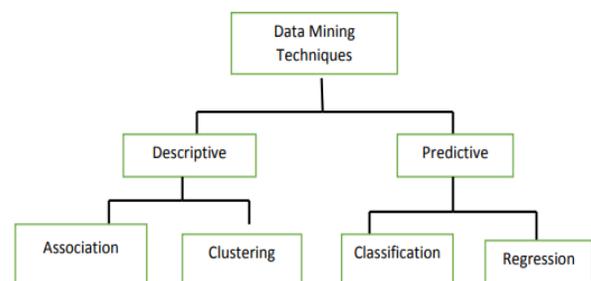


Fig. 1. A Flow Diagram Illustration of Data Mining Techniques.

1) *Decision tree*: Decision trees are used as classifier representations and are constructed using data to solve research problems such that attribute tests are denoted by non-leaf nodes and test outcomes denoted by branches while leaf nodes are assigned particular class levels [8]. Decision tree analysis is used by researchers to determine conditional probabilities for optimal decision making and class separation based on information gain. In the healthcare field, decision trees are used in the classification discrete values due to the ability to process nominal and numeric attributes while adjusting missing data values.

2) *Support Vector Machine (SVM)*: Support vector machine is an advanced classification algorithm for linear and non-linear data sets. It is applied in the transformation of original training data to higher dimensions at which an optimal hyperplane that separates class instances can be determined. Support and marginal vectors provide a framework for determining the hyperplane in SVM subject to the kernel metric $C = J$ [13].

3) *Meta learning classifier*: This is a classification approach in which historical data is used as a learning set using algorithms such as the random subspace, adaboost, and bagging. The adaboost algorithm is applied to improving the classification accuracy by performing multiple iterations to cluster weak learning algorithms and modifying the accuracy parameters, especially in imbalanced or misclassified sets. Adaboost algorithm is implemented as shown in Fig. 2 [14].

The bagging algorithm is implemented through bootstrap aggregation, which involves deriving base classifiers from the decision tree. Bootstrap samples D_1, D_2, \dots, D_n are selected from a data set D provide the base classifiers C_1, C_2, \dots, C_n [15]. Supposing that an optimal number of votes are assigned to a class for randomly selected labels, then the algorithm extracts training object and classifier sets for bootstrapping after which an integration process based on majority voting takes place [16]. The implementation procedure for the bagging algorithm is illustrated in Fig. 3.

Ensemble learning technique describes a process in which multiple classifiers are trained to generate decision insights based on different classifiers through random subspace, bagging, and boosting approaches for increased performance [17]. The most common ensemble learning approaches include weighted averages, majority voting, and simple averages. Ensemble techniques combines multiple classifiers in determining the optimal classification model from different sub-models comprising of a base classifier layer and meta-classifier layers, which make accurate predictions [17].

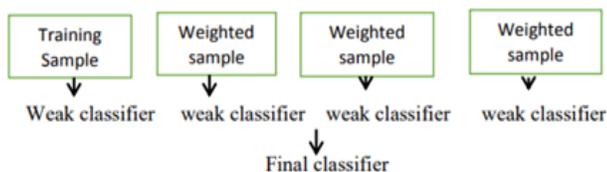


Fig. 2. The Implementation Stages of Adaboost Algorithm.

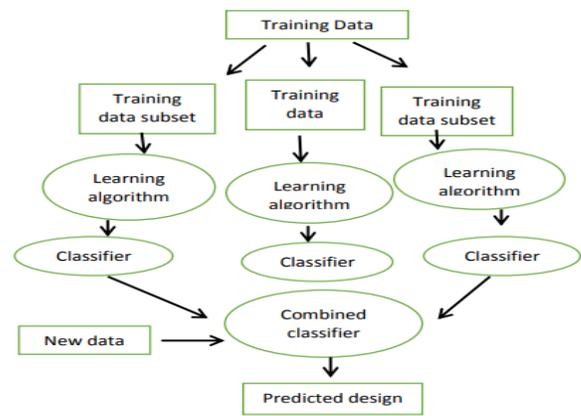


Fig. 3. The Implementation Stages of Bagging Algorithm Ensemble Learning [15].

4) *Attribute subset selection*: Attribute selection techniques play a significant role of data reduction for more efficient analysis in the data mining process. When data sets have many attributes, attribute selection is used to determine those that align to the cost of data analysis and utility for the easier discovery of patterns. Filter and wrapping categorization methods are used in evaluate the estimation accuracy of the learning algorithm [18].

5) *Resampling, oversampling, and under sampling method*: Data mining techniques are applied in healthcare to identify emerging trends from unstructured data sets. Resampling methods combine multiple approaches, which include the Random-oversampling of minor data classes, random oversampling of major classes hence providing solutions to sample distribution problems. Under-sampling removes data imbalances through the random elimination of major classes while oversampling achieves the same objective by replicating minor classes [19].

III. METHODOLOGY

This paper utilizes an integrated methodological framework for literature review, dataset extract, and pre-processing to prepare it for analysis. The primary requirements obtained for this project included the datasets, relevant tools for pre-processing to identify the missing values, models for attribute selection and cross validation, data resembling framework, and intelligent algorithms for base classification. Additionally, the ensemble model and meta-learning algorithms were acquired in preparation for performance evaluation by embedding feature selecting capabilities into the classification model.

A. Dataset Description and Data Transformation

The data used for this study was obtained from the prostate cancer unit at Mayo Clinic, Rochester from a sample population of 1144 patients whose attributes such as age, size of tumor, Node-caps, degree of malignancy, metastasis, and class were recorded. Data imbalances were detected in 808 zero reoccurrences and 336 recurrences as shown in the Table I.

TABLE I. DATASET DESCRIPTION

Attribute	Description	Attribute Type
Tumor	Swollen prostates	Numeric
Age	Age of the patient	Numeric
Node	Absence or presence of node	Nominal
Metastasis	Tumor spread throughout the body	Nominal
Class	Recurrence of risk factors	Nominal
Degree of malignancy	Stage of cancer development	Numeric

WEKA open source software was selected to perform the data mining processes in this study. This tool has integrated data mining capabilities for clustering, regression analysis, classification, pre-processing, and visualization [20]. Pre-processing was performed to ensure that the attribute types of each data class was either nominal or numeric and all missing values replaced with the computed average. Imbalance problems in the dataset were resolved through resampling techniques and the attributes were selected through a dimensionality reduction technique with optimal gain ratio.

B. Classification Algorithms Selection

The classifier algorithms selected for data mining in this study include J48, Neural Networks, Rep Trees, and SVM as the base classifiers, and meta-classifiers such as random subspace, boosting, and bagging, which were used in the building of classifier models.

1) *Decisions tree J48 algorithm*: The basic algorithm involved processes such as the construction of decision trees using the top-down divide-and-conquer approach with root training examples based on the categorical classification of attributes [21]. Recursive partitioning of the heuristic measures was implemented under conditions that all samples are assigned to the same classes and leaf classification based majority voting for all samples [22].

2) *Neural network algorithm*: The data input is embedded simultaneously into input layer after which it is weighted and adopted to a hidden layer, which is usually arbitrary. The last hidden layer contains weighted outputs which form the output layer, which produce predictive insights about the network patterns [20]. A feed-forward approach is applied to the network such that weight cycles in the input or output units are not returned to their previous layer.

3) *Rep tree algorithm*: This algorithm prunes the decision tree to allow the re-generation of the initial tree with minimal error. The data instances are segmented into multiple units such that set leaf can be assigned the lowest number of instances [23].

4) *SVM algorithm*: A relatively new classification method for both linear and nonlinear data, It uses a nonlinear mapping to transform the original training data into a higher dimension [24].

With the new dimension, it searches for the linear optimal separating hyperplane (i.e. “decision boundary”).

With an appropriate nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated by a hyperplane.

SVM finds this hyperplane using support vectors (“essential” training tuples) and margins (defined by the support vectors).

5) *Bagging algorithm*: This algorithm classifies datasets into training and testing categories. Multiple training sets are generated and replaced in iterative sequences to reduce the likelihood of over-fitting and control variance.

6) *Boosting algorithm*: The implementation of this algorithm follows an iterative procedure for adaptive classification of training datasets, especially the misclassified sets. The algorithm assigns equal weights to the initial records N and performs automatic adjustments unit weights are increased in the wrongly classified datasets and increased in the accurately classified data sets [25].

7) *Random subspace algorithm*

Repeat for $b = 1, 2, \dots, B$.

Choose an r -dimensional random subspace b from the original p -dimensional feature space X .

Build a classifier $C_b(x)$ (with a decision boundary $C_b(x) = 0$) in b .

Aggregate classifiers $C_b(x), b = 1, 2, \dots, B$, by majority voting for the final decision. [26].

C. Ensemble Learning

Ensemble model was applied to a combination of classifiers to determine the point at which classification performance is optimal. Ensemble learning model is composed of the base classifier and meta-classifier layers which receive and analyze prediction inputs to generate the desired output.

D. Evaluation Approach and Techniques

The ensemble model is utilized to classify prostate cancer data using combined sub-classifiers to improve performance and accuracy. Factors such as the relative accuracy of measures, degree of training and simulation errors, and classifier performance are used to validate the model [27]. Recall and precision measures are used to determine the accuracy of classification techniques [28]. Additionally, each classifier is evaluated on the basis of computation time matrix, which shows the rate at which algorithms make correct and incorrect predictions compared to the actual values defined in the dataset [29]. The evaluation of metrics accuracy is illustrated in the Table II.

TABLE II. EVALUATION OF CONFUSION METRICS ACCURACY

	Positive Prediction Class	Negative Prediction Class
Real Class Positive	True Positive	False Negative
Real Class Negative	False Positive	True Negative

True Positive: Accurate classification of recurrence instances.

True negative: Inaccurate classification of no-recurrence instances.

False positive: Inaccurate classification of no recurrence instances as recurrent instances.

False negative: Inaccurate classification of recurrence instances no-recurrence instances.

In order to get TP rate, FP rate, Precision, Recall, F-Measure, Accuracy were used in this research as follows:

1) True Positive (TP) rates (sensitivity/recall) – is the proportion of the actual recurrence (or no recurrence) cases correctly classified.

$$TP \text{ (recurrence)} = TP / (TP + FN)$$

$$TP \text{ (no recurrence)} = TN / (TN + FP)$$

2) False Positive (FP) rates (1-specificity/false alarms) – proportion of actual no recurrence (or recurrence) cases misclassified.

$$FP \text{ (recurrence)} = FP / (FP + TN)$$

$$FP \text{ (no recurrence)} = FN / (FN + TP)$$

3) Precision – proportion of predicted recurrence (or no recurrence) cases that were correct classified.

$$\text{Precision (recurrence)} = TP / (TP + FP)$$

$$\text{Precision (no recurrence)} = TN / (FN + TN)$$

4) Recall–Proportion of predicted recurrence (or no recurrence) cases that were correct classified.

$$\text{Recall} = TP / (TP + FN)$$

5) F—one of the performance measures that is used to retrieve data:

$$F\text{-measure} = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) = 2 \times TP / (2 \times TP + FP + FN)$$

6) Accuracy – proportion of the total predictions that was correct.

$$\text{Accuracy} = TP + TN / (TP + TN + FP + FN). [30]$$

E. Receiver Operation Characteristic (ROC) Curve

ROC curves are used in the summarization of classifier performance based on the analysis of error rates involving false positives and true positives. Acceptable performance metrics are defined by the area under curve and represents the optimal decision boundaries for measuring the estimated costs of instance misclassification [31].

IV. EXPERIMENTAL PROCEDURES AND RESULTS

This section discusses the experimental procedures involving base classifiers with selected attributes, without selected attributes and resampling method in the first experiment while the second involved meta classifiers with

selected attributes, without selected attributes and resampling method.

A. Dataset

The dataset used for these experiments was obtained from the prostate cancer department at Mayo Clinic. The instances are defined by the attributes defined in the methodology section. The data was pre-processed and analyzed using WEKA software. The table shown in Fig. 4 shows how the dataset appeared after preparation using the software.

A	B	C	D	E	F	G	H	I	J
age	tumor-size	node-cap	deg-malig	breast	Metastasi	irradiat	Class		
29	3.6	yes		3 R	yes	no	recurrence-events		
49	3	no		1 R	no	no	no-recurrence-events		
53	4.8	no		2 L	yes	no	recurrence-events		
26	1.6	yes		3 R	yes	yes	no-recurrence-events		
40	1.8	yes		2 L	yes	no	recurrence-events		
36	0.8	no		2 R	no	yes	no-recurrence-events		
34	4.2	yes		3 L	no	no	no-recurrence-events		
36	1.9	no		2 L	yes	no	no-recurrence-events		
26	4.8	no		2 R	no	no	no-recurrence-events		
26	1	yes		2 R	yes	yes	no-recurrence-events		
49	1.1	no		2 L	no	no	no-recurrence-events		
52	4.8	no		2 R	yes	no	no-recurrence-events		
59	2.1	no		1 R	yes	no	no-recurrence-events		
26	2.5	yes		2 R	no	no	no-recurrence-events		
53	2.7	no		2 L	yes	yes	recurrence-events		
57	1.4	no		3 L	no	no	no-recurrence-events		
60	1.9	yes		1 R	no	no	no-recurrence-events		
49	2.9	yes		2 R	yes	no	no-recurrence-events		
37	3.5	no		2 L	no	no	no-recurrence-events		
26	2.7	no		3 L	no	no	no-recurrence-events		
34	2.1	no		1 L	no	no	recurrence-events		
26	2.3	no		2 R	no	yes	no-recurrence-events		

Fig. 4. Sample of Data After Preparing.

B. First Experiments and Results

The first experiment involving base classifiers was conducted to investigate the performance of different algorithms involving imbalanced prostate cancer data sets. The algorithms were applied to data through sampling techniques, without attribute selection, and with attribute selection.

1) *Result of support vector machine:* The implementation of SVM algorithm to data classification without attribute selection had a performance accuracy of 70.63% within duration of 0.15 seconds, 0.3 mean absolute error, kappa statistic 0, relative absolute error 70.8%, and 118.99% root relative squared error as shown in the Fig. 5.

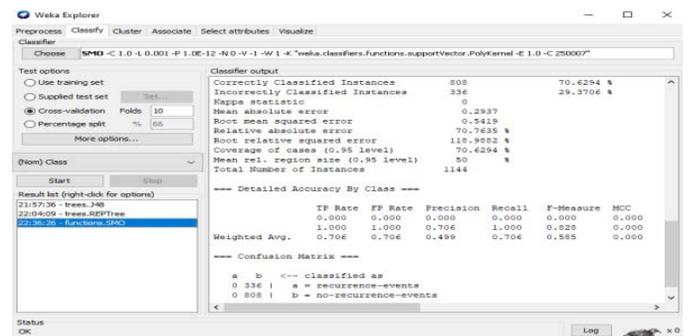


Fig. 5. Result of Classification Model using SVM without Attribute Selection.

The performance outcomes of other base classifier algorithms are shown in the Table III.

TABLE III. RESULTS OF A BASE CLASSIFIER ALGORITHMS WITHOUT ATTRIBUTE SELECTION

Evaluation Criteria	Classifier			
	J48	SVM	ANN	Rep Tree
Duration (seconds)	0.12	0.15	1.54	0.06
Correct classification	799	808	771	792
Incorrect classification	345	336	373	352
Percentage Accuracy	69.8%	70.63%	67.39%	67.39%

The experimental results of algorithm implementation of base classifiers with attribute selection are shown in the Table IV.

TABLE IV. RESULTS OF A BASE CLASSIFIER ALGORITHMS WITH ATTRIBUTE SELECTION

Evaluation Criteria	Classifier (With ranker and gain ratio)			
	J48	SVM	ANN	Rep Tree
Duration (seconds)	0	0.14	1.24	0.06
Correct classification	799	808	771	792
Incorrect classification	345	336	373	352
Percentage Accuracy	69.8%	70.63%	67.39%	69.23%

The performance of experimental parameters was evaluated based on criteria such as the mean errors and kappa statistic is shown in the Table V.

TABLE V. SIMULATION RESULTS

Evaluation Criteria	Classifier (With Ranker and Gain Ratio)			
	J48	SVM	ANN	Rep Tree
Kappa Statistic	0.0038	0	0.035	0.024
Mean Absolute error	0.414	0.294	0.411	0.042
Root mean squared error	0.462	0.542	0.472	0.472
Relative absolute squared error	99.78%	70.76%	99.06%	100.39%
Root relative squared error	101.42%	118.98%	103.63%	103.67%

2) *Experiment using rep tree with resampling method:* Resampling technique was applied on the base classifiers and implemented on decision tree rep to obtain accuracy scores in the data classification. The implementation results for each classifier algorithm are shown the Table VI.

C. Second Experiment and Results

The second experiment involved the analysis of performance classification scores on meta learning algorithms with and without attribute selection. The relative accuracy values are shown in the Table VII and Table VIII.

1) *Evaluation of algorithms:* The algorithms were further evaluated using criteria such as the mean errors and Kappa statistics and the results are shown in the Table IX.

TABLE VI. REP TREE WITH RESAMPLING METHOD

Evaluation Criteria	Classifier (With Resampling)			
	J48	SVM	ANN	Rep Tree
Duration (seconds)	0	0.06	0.93	0.01
Correct classification	799	808	771	792
Incorrect classification	345	336	373	352
Percentage Accuracy	69.84%	70.63%	71.24%	77.27%

TABLE VII. RESULTS OF META CLASSIFIERS WITHOUT ATTRIBUTE SELECTION

Evaluation Criteria	Classifier		
	Bagging	Boosting	Random Subspace
Duration (seconds)	0.58	0.47	0.2
Correct classification	795	808	807
Incorrect classification	349	336	337
Percentage Accuracy	64.9%	70.63%	70.54%

TABLE VIII. RESULTS OF META CLASSIFIERS WITH ATTRIBUTE SELECTION

Evaluation Criteria	Classifier		
	Bagging	Boosting	Random Subspace
Duration (seconds)	0.27	0.17	0.13
Correct classification	795	808	807
Incorrect classification	349	336	337
Percentage Accuracy	64.9%	70.63%	70.54%

TABLE IX. SIMULATION RESULTS

Evaluation Criteria	Classifier		
	Bagging	Boosting	Random Subspace
Kappa Statistic	0.045	0	-0.001
Mean Absolute Error	0.414	0.411	0.413
Root mean squared error	0.4665	0.4555	0.4553
Relative Absolute error	99.7%	98.95%	99.54%
Root relative squared error	102.42%	100.01%	99.96%

V. RESULTS AND DISCUSSION

The experimental results suggest that a combination of boosting and bagging classification algorithms achieve a higher level of accuracy when resampling is applied to SVM and rep tree. Resampling method effectively improve the accuracy of ensemble learning model when applied to imbalanced datasets on prostate cancer [32]. When each the performance of each algorithm is analyzed after resampling, algorithms with single classifiers such as SVM, neural network, rep, and J48 are more accurate and require less computational time. The experimental outcomes of all trials suggest that the implementation of ensemble learning model yields higher classification accuracy

on J48 tree after resampling compared to before resampling imbalanced datasets. The relative performance of each base classifier on the ensemble model under different conditions of resampling is shown in the Fig. 6.

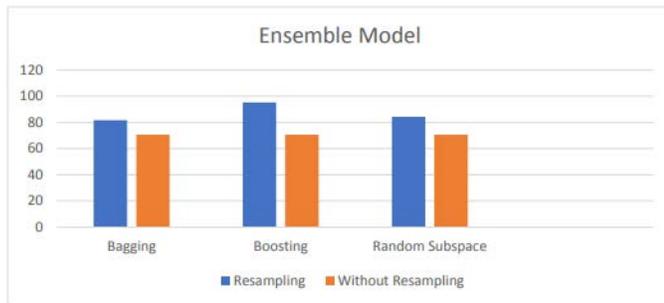


Fig. 6. Comparative Graph for different base Classifiers with different Evaluation Accuracy of Ensemble Model.

VI. CONCLUSION

The objective of this study was to develop a classification model for imbalanced prostate cancer datasets from Mayo Clinic, Rochester. The implementation of accurate classification approaches is important in the early detection and prediction of likelihood of recurrence or no-recurrence of risk factors. The experimental results led to the conclusion that the application of ensemble learning algorithm on resampled data sets provides highly accurate classification results on single classifier J48. The study further suggests that gain ratio and ranker techniques are highly effective for attribute selection in the analysis of prostate cancer data. The lowest error rate and optimal performance accuracy in the classification of imbalanced prostate cancer data is achieved using when Adaboost algorithm is combined with single classifier J48.

The following recommendations were developed based on the empirical results obtained from this study; Consider larger datasets to improve the accuracy of results, implement multiple evaluation techniques, and formulate alternative prediction models and algorithms to allow for the comparative analysis of classification results for imbalanced data.

REFERENCES

- [1] B. Murray, "The Pathogenesis of Prostate Cancer," in Prostate Cancer, Xon Publications, 2021, pp. 29-41.
- [2] S. Saeed and H. C. Ong, "A bi-objective hybrid algorithm for the classification of imbalanced noisy and borderline data sets," Pattern Analysis and Applications, vol. 22, pp. 979-998, 2019.
- [3] A. Elhassan, M. Aljourf, F. Al-Mohanna and M. Shoukri, "Classification of Imbalance Data using Tomek Link (T-Link) Combined with random under-sampling (RUS) as a data reduction method," Global Journal of Technology & Optimization, 2016.
- [4] A. S. Pranto and M. K. Paul, "Performance Analysis of Ensemble Based Approaches to Mitigate Class Imbalance Problem after Applying Normalization," in 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021.
- [5] M. Ye and C. Wu, "Cancer Classification with a Cost-Sensitive Naive Bayes Stacking Ensemble," Computational and Mathematical Methods in Medicine, vol. 2021, pp. 1-12, 2021.
- [6] C. H. Pernar, E. M. Ebot, K. M. Wilson and L. A. Mucci, "The Epidemiology of Prostate Cancer," Cold Spring Harb Perspect Med, vol. 8, no. 12, 2018.
- [7] A. E. Karrar, "The Use of Case-based Reasoning in a Knowledge-based (Learning) Software Development Organizations," International Journal of Innovative Research in Science, Engineering and Technology, vol. 5, no. 5, 2016.
- [8] P. Ahmad, S. Qamar and S. Q. A. Rizvi, "Techniques of Data Mining In Healthcare: A Review," International Journal of Computer Applications, vol. 15, pp. 38-50, 2015.
- [9] N. Padhy, P. Mishra and R. Panigrahi, "The Survey of Data Mining Applications And Feature Scope," International Journal of Computer Science, Engineering and Information Technology, vol. 2, no. 3, pp. 43-58, 2012.
- [10] R. H. Alsagheer, A. F. Alharan and A. S. A. Al-Haboobi, "Popular Decision Tree Algorithms of Data Mining Techniques: A Review," International Journal of Computer Science and Mobile Computing, vol. 6, no. 6, pp. 133-142, 2017.
- [11] M. Mutasim and A. Karrar, "Impute Missing Values in R Language using IBK Classification Algorithm," International Journal of Engineering Science and Computing, vol. 11, no. 6, pp. 28328-28338, 2021.
- [12] A. E. Karrar, "A Novel Approach for Semi Supervised Clustering Algorithm," International Journal of Advanced Trends in Computer Science and Engineering, vol. 6, no. 1, pp. 1-7, 2017.
- [13] M. A. Yaman, A. Subasi and F. Rattay, "Comparison of Random Subspace and Voting Ensemble Machine Learning Methods for Face Recognition," Symmetry, vol. 10, no. 11, 2018.
- [14] L. Zhao, Z. Shang, A. Qin, T. Zhang, L. Zhao, Y. Wei and Y. Y. Tangd, "A cost-sensitive meta-learning classifier: SPFCNN-Miner," Future Generation Computer Systems, vol. 100, pp. 1031-1043, 2019.
- [15] N. Joshi and S. Srivastava, "Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees)," International Journal of Computer Science and Mobile Computing, vol. 3, no. 5, pp. 727-732, 2014.
- [16] D. P. Rangasamy, S. Rajappan, A. Natarajan, R. Ramasamy and D. Vijayakumar, "Variable population-sized particle swarm optimization for highly imbalanced dataset classification," Computational Intelligence, vol. 37, pp. 873-890, 2021.
- [17] M. Mohammed, H. Mwambi, B. Omolo and M. K. Elbashir, "Using stacking ensemble for microarray-based cancer classification," in International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCEEE), 2018.
- [18] T.-B. A.J., C. L. and L.-D. R., "Attribute Subset Selection for Image Recognition. Random Forest Under Assessment," in 16th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2021). SOCO 2021. Advances in Intelligent Systems and Computing., 2022.
- [19] H. Li and M. Zhuang, "Clustering Center Optimization under-Sampling Method for Unbalanced Data," Journal of Software, vol. 15, no. 3, pp. 74-85, 2020.
- [20] J. Nuhic and J. Kevric, "Prostate Cancer Detection Using Different Classification Techniques," in CMBEBIH 2019, IFMBE Proceedings, Springer, Cham., 2020.
- [21] A. M. Poonia, S. Vigneshwari and D. J. Rani, "Machine Learning based Diabetes Prediction using Decision Tree J48," in 2020 3rd International Conference on Intelligent Sustainable Systems, 2020.
- [22] M. F. Maulana and M. Defriani, "Logistic Model Tree and Decision Tree J48 Algorithms for Predicting the Length of Study Period," Journal Penelitian Ilmu Komputer, System Embedded & Logic, vol. 8, no. 1, pp. 39-48, 2020.
- [23] R. Naseem, B. Khan, A. Ahmad, A. Almogren, S. Jabeen, B. Hayat and M. A. Shah, "Investigating Tree Family Machine Learning Techniques for a Predictive System to Unveil Software Defects," Complexity, vol. 2020, no. 6, pp. 1-21, 2020.
- [24] N. Hafidz, Sfenrianto, Y. Pribadi, E. Fitri and Ratino, "ANN and SVM algorithm in Divorce Predictor," International Journal of Engineering and Advanced Technology, vol. 9, no. 3, pp. 2523-2527, 2020.
- [25] M. M. Nishat, T. Hasan, S. M. Nasrullah, F. Faisal, A.-A.-R. Asif and A. Hoque, "Detection of Parkinson's Disease by Employing Boosting Algorithms," in 2021 Joint 10th International Conference on Informatics, Electronics & Vision (ICIEV) and 2021 5th International Conference on Imaging, Vision & Pattern Recognition (icIVPR), Kitakyushu, Japan, 2021.

- [26] M. A. Corsetti and T. M. Love, "Grafted and vanishing random subspaces," in *Pattern Analysis and Applications*, Springer, 2021.
- [27] A. Doganer, "Different Approaches to Reducing Bias in Classification of Medical Data by Ensemble Learning Methods," *International Journal of Big Data and Analytics in Healthcare*, vol. 6, no. 2, pp. 15-30, 2021.
- [28] M. Umair, F. Majeed, M. Shoaib, M. Q. Saleem, M. S. Adrees, A. E. Karrar, S. Khurram, M. Shafiq and J.-G. Choi, "Main Path Analysis to Filter Unbiased Literature," *Intelligent Automation and Soft Computing*, vol. 32, no. 2, pp. 1179-1194, 2022.
- [29] C. N.V., *Data Mining for Imbalanced Datasets: An Overview.*, *Data Mining and Knowledge Discovery Handbook*, Springer, Boston, MA., 2009, pp. 875-886.
- [30] P. Cristaldo, D. D. Luise, L. L. Pietra, A. D. Battista and D. Hemanth, "Data Mining-Based Metrics for the Systematic Evaluation of Software Project Management Methodologies," *EAI/Springer Innovations in Communication and Computing*. Springer, Cham., 2022.
- [31] G. B. Demisse, T. Tadesse and Y. Bayissa, "Data Mining Attribute Selection Approach for Drought Modelling: a Case Study for Greater Horn Of Africa," *International Journal of Data Mining & Knowledge Management Process*, vol. 7, no. 4, pp. 1-16, 2017.
- [32] Y. Wang, D. Wang, N. Geng, Y. Wang, Y. Yin and Y. Jin, "Stacking-based ensemble learning of decision trees for interpretable prostate cancer detection," *Applied Soft Computing*, vol. 77, pp. 1-37, 2019.

Dynamic Deployment of Road Side Units for Reliable Connectivity in Internet of Vehicles

Abdulwahab Ali Almazroi, Muhammad Ahsan Qureshi

University of Jeddah, College of Computing and Information Technology at Khulais
Department of Information Technology
Jeddah, Saudi Arabia

Abstract—Internet of vehicles (IoV) promises to provide ubiquitous information exchange among moving vehicles and reliable connectivity to the internet. Therefore, IoV is becoming more and more popular as the number of connected vehicles is increasing. However, the existing vehicular communication infrastructure cannot guarantee reliable connectivity because all-time information exchange for every travelling vehicle is not assured due to lack of required number of roadside units (RSUs) especially along the intercity highways. This study is aimed towards exploring the use of cost-effective dynamic deployment of RSUs based upon the road traffic density and by ensuring the Line of Sight (LOS) among RSUs and Cellular Network Antennas. The unmanned aerial vehicles (UAVs) have a potential to serve as economical dynamic RSUs. Therefore, the use of UAVs along the roadside for providing reliable and ubiquitous information exchange among vehicles is proposed. The UAVs will be deployed along the roadside and their respective placement will be changed dynamically based upon the current traffic density in order to ensure the all-time connectivity with the travelling vehicles and the other UAVs/Cellular Network antennas. The reliability of the proposed network will be tested in terms of signal strength and packet delivery ratio (PDR) using the simulation.

Keywords—VANET; Roadside unit; internet of vehicle; social internet of vehicles; unmanned aerial vehicle; line of sight vehicular communication

I. INTRODUCTION

Vehicular Ad Hoc Networks (VANETs) [1] provide vehicle to infrastructure (V2I) and vehicle to vehicle (V2V) communication in order to deliver extensive range of applications spanning from infotainment to safety related applications [2, 3]. VANET is gradually evolving into Internet of Vehicles (IoV) that provides internet access to the travelling vehicles in addition to the inter-vehicular communication [4, 5]. The popularity of IoV is increasing rapidly due to its huge potential in terms of connectivity and a wide range of advantageous applications [6]. The modern vehicular communication also offers accident detection [7]. The communication in VANETs depends on the Roadside Unit (RSU) [8]. However, this massive connectivity requires optimal deployment of communication infrastructure; in which roadside units (RSUs) are important component. RSUs provide the essential link among the vehicles and other infrastructure components to fulfill the connectivity challenges of IoV [9]. RSUs connect the vehicles to the other vehicles which are outside their radio range and also with the infrastructure

components in order to provide internet access to the vehicles on the road [10].

A comprehensive strategy for installation of RSUs is proposed in [10]. In [11], RSUs are placed by utilizing geometrical methods to improve the LOS among communicating vehicle in modern road infrastructure units. Another scheme introduces deployment of RSUs efficiently in modern roads using approximation algorithm [12]. A similar approach is used in [13]. Most of the existing studies [14-16] focus on the optimized static placement of RSUs by using innovative algorithms to assist vehicular communication. On the other hand, the recent advancements in unmanned aerial vehicles (UAVs) technology can assist the vehicular communication by solving the problem of ubiquitous internet connectivity and all-time inter-vehicle communication. Therefore, a few recent studies [17, 18] tend to exploit the use of UAVs (drones)[19]as relays to assist vehicular communication where the road traffic is dense by improving the performance and reducing the network delay. UAVs are becoming efficient, lightweight and economical gradually [19, 20]. A few studies advocate the use of UAVs to assist vehicular communication; however, the focus is to improve the performance and reduce the network delay in dense environment [21]. However, the notion of modern IoV requires ubiquitous internet connectivity and all-time inter-vehicle communication especially when traveling on highways where the road traffic density is sparse; therefore, internet connectivity and vehicular communication are compromised especially in highway environment [12]. Therefore, the existing studies lack in the exploitation of UAVs to serve as potential dynamic RSU especially for the highway (sparse) environment as the RSUs have limited converge area that results in significant degradation in communication. Currently, static placement of RSUs is in practice. The placement of a RSU can be changed accordingly with change in the road traffic density by ensuring the Line of Sight (LOS) among RSUs and Cellular Network Antennas to overcome the limited coverage area issue of RSUs.

The goal of this study is to address the problem of all-time inter-vehicle communication and ubiquitous internet connectivity. Therefore, the aim of current work is to investigate the use of cost-effective dynamic deployment of RSUs based upon the road traffic density and by ensuring the Line of Sight (LOS) among RSUs and Cellular Network Antennas. The study targets the UAVs for economical dynamic deployment of RSUs due to their efficient, lightweight and

economical nature. Therefore, the use of UAVs along the roadside for providing reliable and ubiquitous information exchange among vehicles is proposed. Consequently, the current study proposes dynamic deployment and placement of UAVs along the roadside based upon the current traffic density. As a result, the proposed method ensures all-time connectivity with the travelling vehicles and the other UAVs/Cellular Network antennas.

To achieve the goal of the study, the objectives of the research are as follows:

Objective 1: To study state-of-the-art in the domain of deployment of RSUs.

Objective 2: To propose dynamic deployment of RSUs based upon road traffic density using UAVs.

Objective 3: To evaluate the proposed solution in terms of connectivity index.

Objective 4: To compare the proposed framework with and without dynamic deployments of RSUs existing solution(s).

The current study follows a research design consisting of five main steps to achieve the objectives of the work. In the first step, a review of state-of-the-art studies in roadside unit deployment is performed. Step 2 focuses on comparison of existing energy efficiency techniques. Developing a cost effective dynamic RSU deployment strategy based on efficient utilization of UAVs is the target of step 3. Evaluation of the proposed solution and comparison of proposed solution with the existing solutions are part of step 4 and 5.

The rest of the paper is organized as follows: Section 2 presents state-of-the-art studies in roadside unit deployment. This is then followed by Section 3, which describes cost effective dynamic RSU deployment strategy. The evaluation of proposed strategy is explained in Section 4. Section 5 provides a comparison between proposed strategy and existing work. The research implications of the study are highlighted in Section 6 and finally, Section 7 concludes the paper.

II. RELATED WORK

Multiple studies were conducted in the past for the efficient deployment of RSUs in VANETs and IoVs. This subsection reviews a few important studies in the domain of RSUs deployment.

Some research utilized artificial intelligence technology to address various problems of IoV such as energy management, traffic monitoring and management, resource management, big data processing, and communication problem [8, 22, 23]. In [24], the authors used reinforcement learning approach called centralized Q-learning for energy efficiency and optimization in IoV. Similarly, [25]fuzzy quality of service is utilized for optimization of energy in IoV. The author in [22] focused on multi-media communication in IoV. On the other hand, UAVs technology has shown enormous potential to provide efficient solutions to diverse problems belonging to different fields of life due to its cost effective, lightweight and efficient nature. Currently, UAVs have wide range of applications in different domains like smart cities, traffic management, military, smart agriculture, smart healthcare, smart houses and industry [26-

29]. In smart cities, UAVs are providing many services such as traffic management, environmental monitoring, pollution monitoring, and security control [27]. The solutions provided by UAVs to traffic management include [28] that proposes a system for smart traffic monitoring to overcome limitations of existing system. Due to the recent advancements in UAVs and the cost-effective nature of UAVs, the study aims to utilize UAVs for providing all-time inter-vehicle communication and ubiquitous internet connectivity by dynamic deployment of RSUs based upon the road traffic density and by ensuring the Line of Sight (LOS) among RSUs and Cellular Network Antennas.

A mechanism to maintain line of sight (LOS) among the travelling vehicles was proposed in [11]. The study considered modern road infrastructures such as flyovers, underpass, curved roads and tunnels. The study utilizes geometrical concepts for the efficient deployment of RSUs, signal enhancers and signal reflectors to maintain LOS among the travelling vehicles. The results of the study were promising as this concept provides reliable connectivity due to the maintenance of LOS among travelling vehicles. However, this study does not consider the most important highway environment where the connectivity is compromised due to coarse density of vehicles.

A cooperative architecture for Intelligent Transportation System (ITS) [30] based upon distributed RSUs was presented in [31]. Using this architecture, the data from all the sensors was collected in a distributed fashion without the intervention of a central control system. It defines the role of each individual network element in the controlling the sensors and disseminating the information. Real-world experiments were also conducted to prove the correctness of idea. However, this study does not focus on the reliability issues of information dissemination as no discussion is presented on the efficient deployment of RSUs.

In an attempt to identify the optimal number of RSUs in a highway environment, a research was conducted that analyzed the delay of message dissemination in VANETs [32]. Based upon the analysis result, the RSUs were deployed at optimal distances from each other. The experimental results also verify optimal RSUs placement. However, this study does not consider the dynamic RSU deployment in VANET environment.

Another attempt was carried out in order to optimally deploy the RSUs using a novel concept of Minimal Mobility Pattern Coverage (MPC) [14]. Firstly, the mobility of travelling vehicle is predicted from a trace file. Secondly, based upon the extracted information, the optimal placement of RSUs is advocated by extracting minimal traversal of a hyper-graph. The authors claim that the experiment results validate the proposed research. However, this study is yet not verified at large scale. Furthermore, the deployment of RSUs is based upon predicted mobility pattern. Therefore any change in the mobility pattern will negatively affect the optimal placement of RSUs.

In a relatively recent study [33], an algorithmic approach towards optimal placement of one-dimensional RSU is presented. A strategy called "dynamic limiting" is first used for

pruning the search space. A greedy algorithm named “OptDynLim” is proposed that optimally identifies the placement of RSU. The results of the study are formally verified and validated with the help of simulations. However, this study only considers signal dimension RSU deployment problem and don not focuses on general RSU deployment.

Another study [34] exists in the literature that focuses collectively on the 2-D RSU deployment and service task assignment in the domain of IoV. A linear programming based clustering algorithm was proposed that considers the delay requirements and the task assignment. The comparison of the algorithmic output with the optimal solution proved the workability of the proposed solution. However, this study lacks in considering the changing traffic density requirement and dynamic deployment of RSUs.

A study [35] presented a state-of-the-art review of the applicability of UAVs in modern transportation system in smart city environment. This study has highlighted the potential uses of UAVs in ITS and the challenges that may encountered during the implementation. The potential applications include the accident reporting and police eye, along with the flying road side unit. Therefore, this study can also serve as a pivotal point towards the dynamic RSU deployment.

III. PROPOSED FRAMEWORK

In this section, the proposed solution is discussed that provide efficient, light weight, and cost effective method to provide all-time inter-vehicle communication and ubiquitous internet connectivity by dynamic deployment of RSUs based on UAVs. To achieve the objective of the work, first a mathematical model is formulated, then based on the proposed mathematical model, an extensive simulation is carried out to obtained the result. The proposed solution place RSUs dynamically based upon the road traffic density and by ensuring the Line of Sight (LOS) among RSUs and Cellular Network Antennas. In the proposed solution, a road is divided into n segments and a road segment has a specific number of poles. Mostly, these poles are part of infrastructure. One UAV is assigned to every road segment. Based on the traffic density, the UAVs containing RSU can change its location dynamically from one pole to another in the same road segment. This solution is mathematically represented below:

The mathematical model of proposed solution is described below:

$$\text{Let } t \in P \text{ where } p = \{p_1, p_2 \dots p_n\} \tag{1}$$

Here p represents poles in a road segment R.

$$\text{Let } D = \{d_1, d_2, \dots d_m\} \text{ where } m < n \tag{2}$$

Here, D describes the number of drones. It is to be noted that the number of drones is less than the number of poles in a road segment R.

$$R = \{(p_1, p_2), (p_2, p_3), \dots (p_{n-1}, p_n)\} \tag{3}$$

C_i represents number of vehicles in any R at time T_i and d_T is the density threshold. For no change in the scenario, $C_i < d_T$

$$D = \forall d \in D \text{ there exists } \exists p \in P \mid (d, p) \tag{4}$$

Otherwise,

$$D = \forall d_i \in D \text{ there exists } \exists P_i \in P \mid (d_i, p_{i+1}) C_i \geq d_T \tag{5}$$

At time T_1 , D may have following values depicting that which particular drone is present at which pole at the moment.

$$D = \{(d_1, p_1), (d_2, p_4), \dots (d_m, p_{n-1})\} \tag{6}$$

IV. EXPERIMENTAL SETUP

In the current research, three scenarios are selected for performing experimental results. The details of these scenarios are presented in Table I. For the scenario 1 and 2, M2 motorway of Pakistan, and Makkah-Madinah highway are selected, respectively. Khulais, urban area of Makkah province of Saudi Arabia is selected for urban scenario. The topologies of roads selected in our scenarios are imported to SUMO simulator by utilizing OpenStreetMap utility. A total of three dynamic RSUs are deployed with a range of one Km. The minimum and maximum vehicles for highway scenario 1 and 2 are same, which are, 1 and 50, respectively. However, due the different nature of urban scenario the minimum number of vehicles is 10 and the maximum number of vehicles is 60. Further, minimum and maximum communicating vehicles are same for highway scenario while urban scenario has different minimum and maximum vehicles as depicted in Table I. For both highway scenarios a three KM segment of road is selected. On the other hand, nine square KM segment is preferred for urban scenario. Fig. 1 depicted the highway scenario 1, i.e. M2 motorway of Pakistan. Urban scenario in SUMO is portrayed in Fig. 2 and Fig. 3 shows the urban scenario with running vehicles in SUMO.

TABLE I. DETAIL OF EXPERIMENTAL SETUP

Scenarios	Minimum Vehicles	Maximum Vehicles	Minimum Communicating Vehicles	Maximum Communicating Vehicles	Area	Selected Highways
Highway Scenario 1	1	50	1	14	3 KM Segment	M2 Motorway, Pakistan
Highway Scenario 2	1	50	1	14	3 KM Segment	Makkah-Madinah Highway, Saudi Arabia
Urban Scenario	10	60	4	17	9 Square KM	Khulais, Makkah, Saudi Arabia

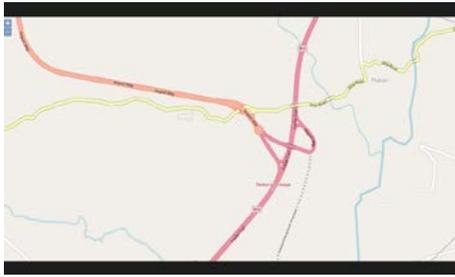


Fig. 1. High Way Scenario.

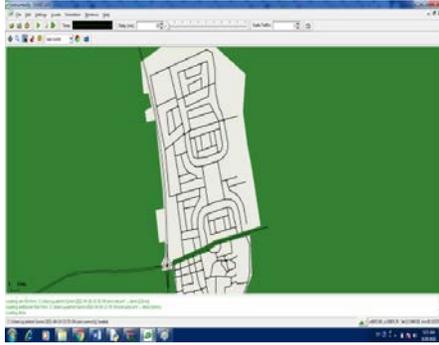


Fig. 2. Urban Scenario in SUMO.

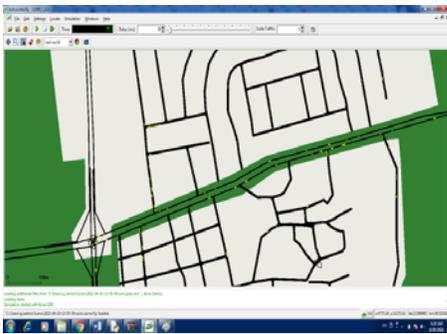


Fig. 3. Urban Scenario with Running Vehicles.

V. EXPERIMENTAL RESULTS

This section presents the results of connectivity index and percentage index for three selected scenarios.

Fig. 4, 5 and 6 described the connectivity index with dynamic deployment of RSUs for scenario 1, 2, and 3 respectively. The connectivity index elevates with the increased number of vehicles, however, important point to note here is, due to dynamic deployment, the connectivity index has raised 19% and 25%, respectively. Consequently, it is inferred that, with the deployment of RSUs dynamically, significant improvement in connectivity index is achieved resulting in enhanced vehicular communication. Correspondingly, connectivity index upgraded 13% in urban scenario supporting the previous inference. However, the boost in vehicular communication in urban scenario is less as compared to highway scenarios. Hence, it is concluded that the proposed dynamic deployment is more suitable to the highway scenarios as compared to urban scenario. Further, extensive research is required for the dynamic deployment of RSUs in urban scenario to take maximum advantage of the proposed deployment.

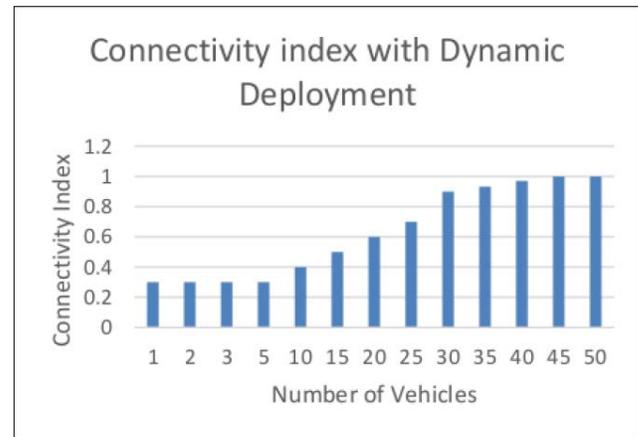


Fig. 4. Connectivity Index for Highway Scenario 1.

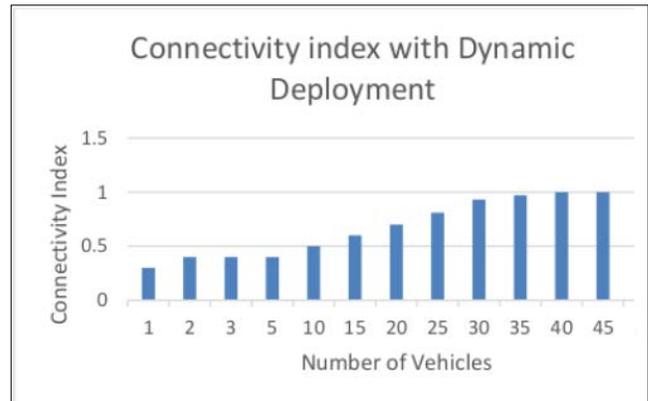


Fig. 5. Connectivity Index for Highway Scenario 2.

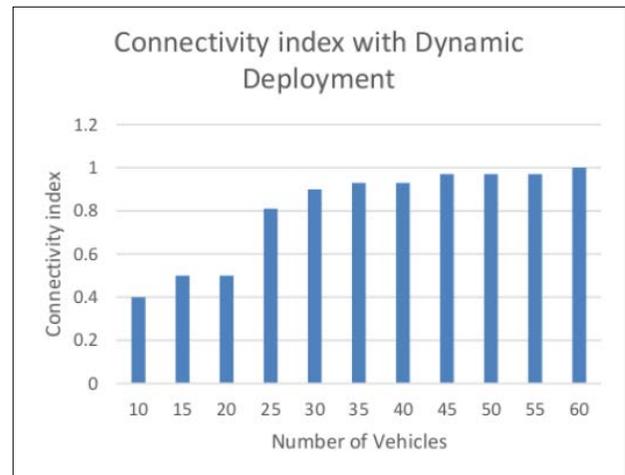


Fig. 6. Connectivity Index for Urban Scenario.

Fig. 7 exhibited the amplification in vehicular communication due to dynamic deployments of RSUs. A considerable escalation is attained in the all scenarios. For highway scenario 1 and 2, gain in connectivity is 19% and 25%, respectively. On the other hand, the gain connectivity is 13%. These results support the dynamic deployment of RSUs in highway scenarios as well as in urban scenarios. However, in urban scenarios, the connectivity gain can be enhanced by conducting further research as discussed earlier.

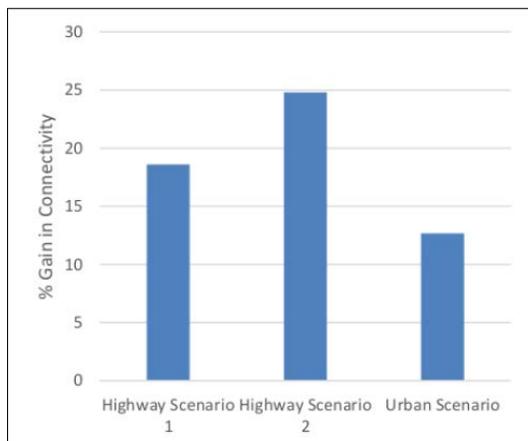


Fig. 7. % Gain in Connectivity of all Scenarios.

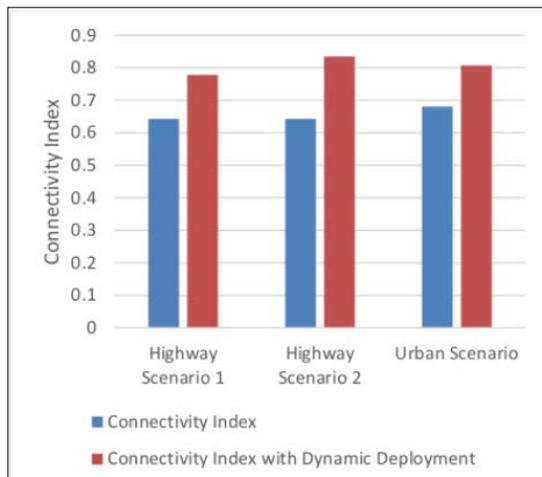


Fig. 8. Comparison of Connectivity Index with and without Dynamic Deployment for all Scenarios.

It is also important to validate the results to show the improvement in connectivity index due the proposed dynamic deployment of RSUs. For this purpose, the current work compares the connectivity index with and without dynamic deployment of RSUs for all scenarios. Fig. 8 illustrates this comparison. A superior connectivity index has achieved in all scenarios due to the deployment of RSUs dynamically. The highway scenario 2 obtained the highest connectivity index while the lowest connectivity index is associated with highway scenario 1 with the proposed method.

VI. RESEARCH IMPLICATIONS

The findings of the studies can be beneficial to for scientific community. The findings can be utilized to design and develop new standards, applications, and protocols on the basis of the dynamic RSU deployment. It will provide a cost effective solution to the classic RSU deployment problem. Further, the implications of the findings of the study are twofold in scientific community perceptive; (i) the new standards, applications, and protocols in IoV and modern transportation domains will be evaluated considering the maximum reliability in message dissemination, and (ii) the cost deployment and maintenance of RSUs will be reduced. In the society betterment prospective, the results can be exploited in

multiple manners. In line with the vision 2030 of Saudi Arabia, the cost effective solution in modern transportation environment along the highways will be beneficial for the general public and transportation companies in terms of reliability of information exchange during the travel. Specially, the expected outcomes are advantageous for huge traffic movement during the Hajj and Umrah seasons.

VII. CONCLUSION, FUTURE WORK AND LIMITATIONS

Currently, IoV provides Internet connectivity between moving vehicles; however, all-time connectivity among moving vehicles is not assured due to the limited number of RSUs. Consequently, busy intercity highways are facing a problem of limited connectivity to the Internet resulting in reduced information exchange. To address this limitation, the current study proposes, mathematically modeled, experimented, and evaluated a scheme to dynamically place RSUs in highways and urban scenarios. The placement of RSUs is based on road traffic density by ensuring line of sight among RSU and Cellular Network Antennas. The proposed scheme is evaluated based on connectivity index that describes the communication among vehicles. The experimental results for three scenarios: highway scenario 1, highway scenario 2, and urban scenario highlighted the superiority of proposed scheme in terms of connectivity index. The connectivity index is considerably higher in the proposed scheme for all of the scenarios resulting in increased vehicular communication. Thus, addressing the problem of vehicular communication in IoV. The study has some limitations too. First, the proposed scheme is evaluated on small scale with three scenarios. Second, for the large scale evaluation, only simulation results are presented. Third, real-world evaluation is not executed. Therefore, future directions of this work include real-world evaluation of the scheme. Moreover, more scenarios from different geographical region will be select to evaluate the proposed scheme.

ACKNOWLEDGMENT

This work was funded by the Deanship of Scientific Research (DSR), University of Jeddah, Jeddah, under grant No. (UJ-20-093-DR)The authors, therefore, acknowledge with thanks DSR technical and financial support.

REFERENCES

- [1] M. A. Qureshi, R. M. Noor, S. Shamshirband, S. Parveen, M. Shiraz, and A. Gani, "A survey on obstacle modeling patterns in radio propagation models for vehicular ad hoc networks," *Arabian Journal for Science and Engineering*, vol. 40, pp. 1385-1407, 2015.
- [2] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 3163-3173, 2019.
- [3] M. Ahsan Qureshi, E. Mostajeran, R. M. Noor, A. Shamim, and C.-H. Ke, "A Computationally Inexpensive Radio Propagation Model for Vehicular Communication on Flyovers and Inside Underpasses," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 10, pp. 4123-4144, 2016.
- [4] O. Kaiwartya, A. H. Abdullah, Y. Cao, A. Altameem, M. Prasad, C.-T. Lin, et al., "Internet of vehicles: Motivation, layered architecture, network model, challenges, and future aspects," *IEEE Access*, vol. 4, pp. 5356-5373, 2016.
- [5] A. A. Almazroi and M. A. Qureshi, "An Energy-aware Facilitation Framework for Scalable Social Internet of Vehicles," *International Journal of Advanced Computer Science and Applications*, 2021.

- [6] J. Contreras-Castillo, S. Zeadally, and J. A. Guerrero-Ibañez, "Internet of vehicles: architecture, protocols, and security," *IEEE Internet of Things Journal*, vol. 5, pp. 3701-3709, 2017.
- [7] M. H. Alkinani, A. A. Almazroi, N. Z. Jhanjhi, and N. A. Khan, "5G and IoT Based Reporting and Accident Detection (RAD) System to Deliver First Aid Box Using Unmanned Aerial Vehicle," *Sensors (Basel, Switzerland)*, vol. 21, 2021.
- [8] E. S. Ali, M. K. Hasan, R. Hassan, R. A. Saeed, M. B. Hassan, S. Islam, et al., "Machine Learning Technologies for Secure Vehicular Communication in Internet of Vehicles: Recent Advances and Applications," *Security and Communication Networks*, vol. 2021, 2021.
- [9] A. Ali, N. Ayub, M. Shiraz, N. Ullah, A. Gani, and M. A. Qureshi, "Traffic Efficiency Models for Urban Traffic Management Using Mobile Crowd Sensing: A Survey," *Sustainability*, vol. 13, p. 13068, 2021.
- [10] D. Kim, Y. Velasco, W. Wang, R. Uma, R. Hussain, and S. Lee, "A new comprehensive RSU installation strategy for cost-efficient VANET deployment," *IEEE Transactions on Vehicular Technology*, vol. 66, pp. 4200-4211, 2016.
- [11] M. A. Qureshi and R. M. Noor, "Towards improving vehicular communication in modern vehicular environment," in *2013 11th International Conference on Frontiers of Information Technology*, 2013, pp. 177-182.
- [12] M. Ge and Y. Chung, "Efficient Deployment of RSUs in Smart Highway Environment," *International journal of advanced smart convergence*, vol. 8, pp. 179-187, 2019.
- [13] Z. Wang, J. Zheng, Y. Wu, and N. Mitton, "A centrality-based RSU deployment approach for vehicular ad hoc networks," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1-5.
- [14] T. Yeferny and S. Allani, "Mpc: A rsus deployment strategy for vanet," *International Journal of Communication Systems*, vol. 31, p. e3712, 2018.
- [15] Z. Gao, D. Chen, S. Cai, and H.-C. Wu, "Optimal and greedy algorithms for the one-dimensional rsu deployment problem with new model," *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 7643-7657, 2018.
- [16] M. F. Faraj, J. F. Sarubbi, C. M. Silva, and F. V. Martins, "A Memetic Algorithm Approach to Deploy RSUs Based on the Gamma Deployment Metric," in *2018 IEEE Congress on Evolutionary Computation (CEC)*, 2018, pp. 1-8.
- [17] Z. Shafiq, R. Abbas, M. H. Zafar, and M. Basher, "Analysis and Evaluation of Random Access Transmission for UAV-Assisted Vehicular-to-Infrastructure Communications," *IEEE Access*, vol. 7, pp. 12427-12440, 2019.
- [18] M. Khabbaz, J. Antoun, and C. Assi, "Modeling and Performance Analysis of UAV-Assisted Vehicular Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, pp. 8384-8396, 2019.
- [19] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Transactions on Wireless Communications*, vol. 16, pp. 3747-3760, 2017.
- [20] G. Jang, J. Kim, J.-K. Yu, H.-J. Kim, Y. Kim, D.-W. Kim, et al., "Cost-Effective Unmanned Aerial Vehicle (UAV) Platform for Field Plant Breeding Application," *Remote Sensing*, vol. 12, p. 998, 2020.
- [21] N. A. Khan, N. Z. Jhanjhi, S. N. Brohi, A. A. Almazroi, and A. A. Almazroi, "A Secure Communication Protocol for Unmanned Aerial Vehicles," *Cmc-computers Materials & Continua*, vol. 70, pp. 601-618, 2022.
- [22] A. H. Sodhro, Z. Luo, G. H. Sodhro, M. Muzamal, J. J. Rodrigues, and V. H. C. De Albuquerque, "Artificial Intelligence based QoS optimization for multimedia communication in IoV systems," *Future Generation Computer Systems*, vol. 95, pp. 667-680, 2019.
- [23] M. B. Hassan, E. S. Ali, R. A. Mokhtar, R. A. Saeed, and B. S. Chaudhari, "NB-IoT: concepts, applications, and deployment challenges," in *LPWAN Technologies for IoT and M2M Applications*, ed: Elsevier, 2020, pp. 119-144.
- [24] H. Park and Y. Lim, "Reinforcement learning for energy optimization with 5g communications in vehicular social networks," *Sensors*, vol. 20, p. 2361, 2020.
- [25] S. Hu, H. Fan, Z. Wang, and Z. Cai, "A Fuzzy QoS Optimization Method with Energy Efficiency for the Internet of Vehicles," *Advances in Networks*, vol. 4, p. 34, 2016.
- [26] P. K. R. Maddikunta, S. Hakak, M. Alazab, S. Bhattacharya, T. R. Gadekallu, W. Z. Khan, et al., "Unmanned aerial vehicles in smart agriculture: Applications, requirements, and challenges," *IEEE Sensors Journal*, 2021.
- [27] N. Mohamed, J. Al-Jaroodi, I. Jawhar, A. Idries, and F. Mohammed, "Unmanned aerial vehicles applications in future smart cities," *Technological Forecasting and Social Change*, vol. 153, p. 119293, 2020.
- [28] N. A. Khan, N. Jhanjhi, S. N. Brohi, R. S. A. Usmani, and A. Nayyar, "Smart traffic monitoring system using unmanned aerial vehicles (UAVs)," *Computer Communications*, vol. 157, pp. 434-443, 2020.
- [29] F. Syed, S. K. Gupta, S. Hamood Alsamhi, M. Rashid, and X. Liu, "A survey on recent optimal techniques for securing unmanned aerial vehicles applications," *Transactions on Emerging Telecommunications Technologies*, p. e4133, 2020.
- [30] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Vehicular Technology Magazine*, vol. 5, pp. 77-84, 2010.
- [31] A. Moreno, E. Osaba, E. Onieva, A. Perallos, G. Iovino, and P. Fernández, "Design and Field Experimentation of a Cooperative ITS Architecture Based on Distributed RSUs," *Sensors*, vol. 16, p. 1147, 2016.
- [32] C. Liu, H. Huang, and H. Du, "Optimal RSUs deployment with delay bound along highways in VANET," *Journal of Combinatorial Optimization*, vol. 33, pp. 1168-1182, 2017.
- [33] Z. Gao, D. Chen, S. Cai, and H.-C. Wu, "Optdynlim: An optimal algorithm for the one-dimensional rsu deployment problem with nonuniform profit density," *IEEE Transactions on Industrial Informatics*, vol. 15, pp. 1052-1061, 2018.
- [34] Y. Ni, J. He, L. Cai, J. Pan, and Y. Bo, "Joint Roadside Unit Deployment and Service Task Assignment for Internet of Vehicles (IoV)," *IEEE Internet of Things Journal*, vol. 6, pp. 3271-3283, 2018.
- [35] H. Menouar, I. Guvenc, K. Akkaya, A. S. Uluagac, A. Kadri, and A. Tuncer, "UAV-enabled intelligent transportation systems for the smart city: Applications and challenges," *IEEE Communications Magazine*, vol. 55, pp. 22-28, 2017.

Customer Satisfaction with Digital Wallet Services: An Analysis of Security Factors

Dewan Ahmed Muhtasim¹, Siok Yee Tan², Md Arif Hassan³, Monirul Islam Pavel⁴, Samiha Susmit⁵

Center for Artificial Intelligence Technology, Faculty of Information Science and Technology

Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia^{1,2,4}

Center for Cyber Security, Faculty of Information Science and Technology

Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia³

Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia⁵

Abstract—This study aimed to determine an efficient framework that caters to the security and consumer satisfaction for digital wallet systems. A quantitative online survey was carried out to test whether the six factors (i.e., transaction speed, authentication, encryption mechanisms, software performance, privacy details, and information provided) positively or negatively impact customer satisfaction. This questionnaire was divided into two sections: the respondents' demographic data and a survey on security factors that influence customer satisfaction. The questionnaires were distributed to the National University of Malaysia's professors and students. A sample of 300 respondents undertook the survey. The survey results suggested that many respondents agreed that the stated security factors influenced their satisfaction when using digital wallets. Previous studies indicated that financial security, privacy, system security, cybercrime, and trust impact online purchase intention. The proposed framework in this research explicitly covers the security factors of the digital wallet. This study may help digital wallet providers understand the customer's perspective on digital wallet security aspects, therefore motivating providers to implement appropriately designed regulations that will attract customers to utilize digital wallet services. Formulating appropriate security regulations will generate long-term value, leading to greater digital wallet adoption rates.

Keywords—Cashless transaction; electronic payment; internet security; consumer satisfaction; e-commerce

I. INTRODUCTION

A digital wallet platform allows individuals to perform electronic transactions using mobile devices, such as smartphones, computers, and other supporting devices. Digital wallets have multiple usages, including making purchases from websites or using mobile devices to make in-store transactions [1]. The current economy is transforming into a cashless economy where daily transactions are being performed using the digital wallet. Several e-commerce companies have developed digital wallet payment systems. The development of digitization through the internet has advanced the transition of globalization and payment systems from manual to online transactions. People have become more reliant on electronic money to perform transactions as a result [2].

The Malaysian government encourages people to go cashless and use digital wallets. A survey by Visa found that nearly half of Malaysians can live without cash. This study illustrated Malaysia's potential to evolve into a cashless society

[3]. The current advanced technologies make it easier for Malaysia to shift toward a cashless society. In support of Industry 4.0, Malaysia has taken significant initiatives to establish the status of a cashless society. According to FinTech News Malaysia's report, approximately 17% of the corporations in the Malaysian financial industry formed a sector to establish digital wallets [4]. Cashless transactions result in enhanced capability, transparency, and accountability [3].

Nevertheless, cybercrime cases still occur, and cases of internet security violations are frequent. Moreover, hackers and tricksters are prone to take advantage of such situations to damage companies and consumers [5]. Thus, customer satisfaction is greatly influenced by their perceptions of security and trust [6]. Customers in the e-banking industry are more advanced, knowledgeable, and demanding [7]. Companies and customers avoid e-commerce operations for several reasons, and security is one of the key reasons [8]. Therefore, security is a serious concern when performing financial transactions through digital transaction methods [9]. Moreover, electronic payment systems are currently facing many difficult hindrances posed by the internet. In comparison, the challenges faced by electronic payment systems are more complicated than the majority of internet security issues [10].

This work was supported by the Ministry of Higher Education Malaysia (FRGS/1/2018/ICT01/UKM/02/5) and Universiti Kebangsaan Malaysia (GUP-2020-060).

Financial services are extending their sector to include smartphones as electronic payment devices. Smartphones are popular banking, payment, budgeting, shopping, and stock trading applications for customers [11]. Due to the growth of the e-commerce industry, the electronic payment platform is becoming increasingly prominent and essential for smartphone users [12]. Smartphones and other electronic devices contain confidential information of their users, including transaction details and passwords. Consumers do not need to carry their wallets or purse while using digital wallets. However, they need to carry at least one electronic device, such as a smartphone or tablet. In the case of a stolen or lost device, the user risks losing personal and confidential information [13].

The digital wallet is still at its infancy stage in Malaysia as it is a newly introduced payment system. Electronic transactions are a safer mode for consumer payment, enabling

sellers to enhance their productivity and increase profits [14]. Electronic transactions are traceable, whereas cash transactions are not traceable. Hence, the electronic payment system is a secure way of performing transactions. However, according to the latest VMware Banking User, a 2020 study shows that nearly half or 46% of Malaysian consumers are uncertain of digital wallet protection and payment applications [15]. Consumers are concerned about privacy and security threats due to the fear of data fraud and spam [16]. Security and privacy impose a significant and positive impact on behavioral intention when using digital payment services [17]. Therefore, e-retailers should concentrate on improving consumers' protection, loyalty, and purchase intent. They should also be attentive to enhanced security when developing a consumer privacy policy [18]. Hence, the security factors have a significant impact on consumer satisfaction toward digital in Malaysia. Therefore, security factors that can affect Malaysian customer satisfaction are required to be evaluated thoroughly. Ali et al [19]. The author in proposed a framework for the security factors that influence consumers in online shopping in Malaysia. It contains five security factors, financial security, privacy, system security, cybercrime, trust, and customer satisfaction. However, the research failed to identify the specific security factors of the digital wallet.

Nevertheless, limited study has been undertaken to identify the specific digital wallet security factors affecting customer satisfaction in Malaysia. This research proposes a framework for security factors that influence consumer satisfaction in Malaysian digital wallet platforms. This model consists of six security factors: transaction speed, authentication, encryption mechanisms, software performance, privacy details, and information provided. This model is created after studying numerous previous studies from similar fields [19], [20], [29], [30], [21]–[28].

This research introduces security factors to the digital wallet industry in Malaysia that impact customer satisfaction. Besides, this research can assist policymakers of digital wallet applications to concentrate on key security factors and improve platform security while developing security regulations, resulting in a higher rate of consumer adoption. In addition, the research also attempted to study students' and professors' usage and understanding of Malaysia's available digital wallet platforms. The respondents' demographics, comprising both genders, were the students and professors in the National University of Malaysia. They have used the digital wallet services at least once for online purchases. Students between the ages of 18 to 25 were the majority of the respondents.

This research aimed to discover the numerous security concerns that may arise while using digital wallet services for electronic payments besides investigating whether the proposed security factors (i.e., transaction speed, authentication, encryption mechanisms, software performance, privacy details, and information provided) affect the consumer's decision when choosing digital wallet services. This study also aimed to identify variables that have a higher effect on customer satisfaction.

This paper is divided into seven sections. Section 1 presents the introduction and scope of this study, whereas Section 2

reviews previous research on similar topics. In Section 3, the framework and hypothesis of this study are discussed. The methods and analysis techniques of the study are presented in Section 4. Findings of the analysis, the study's limitations, and the conclusions are discussed in Section 5, Section 6, and Section 7, respectively.

II. LITERATURE REVIEW

Digital wallet systems are developed to aid various functionalities. The functions of digital wallet systems are categorized into open digital wallets, semi-closed digital wallets, and closed digital wallet types. Mobile and wireless networking technology, such as smartphones, personal digital assistants (PDAs), and laptops, have eased customers' convenience in utilizing such devices to shop virtually via electronic transaction methods. Transactions are more accessible and transparent through this new method [31]. Nevertheless, several variables may impact the satisfaction of consumers toward electronic transactions. Security, privacy, confidence, and consistency significantly affect e-commerce consumers, among other factors [20]. Mobile perceived security risks determine the consumers' perception of security against conducting mobile transactions, especially the risk of losing important information, resulting in financial losses [32]. Customers' willingness to utilize mobile payment services is influenced by the ease of use, comparative advantage, clarity, and perceived protection.

Furthermore, prevalence and observability positively affect an individual's perception of security, whereas concerns about privacy threats negatively influence the perception [33]. Moreover, the continuous technological development and efforts to promote them are reasons for perceived security as one of the intentions to use digital wallets. Consumers would feel safer when using mobile payment if tools are available to protect the payment systems in unexpected incidences [34]. In an empirical study multiple variables were considered when studying consumer's perceived risk and their attitude toward online shopping in Malaysia. The researchers found that the consumer perceived risks negatively affect the consumer attitude, which positively and significantly affects online shopping behavior [35]. Content quality, peer influence, KOL influence, perceived interaction, effort expectation, and perceived trust all substantially impact users' intention to pay, and their tendency has an indirect impact on users' paying behavior [36].

Although the young generation's actual use of digital payments is driven by behavioral intention and promotional strategies, perceived risks are shown to have a negative effect [37]. Another empirical study asserted that customers were reluctant to use digital wallets when they assumed high perceived risk [11]. Moreover, the study also indicated that financial risks are one of the leading consumer perceived risks. Financial risks emerged as the most significant influence that adversely affects consumer attitude. One of the proposed ideas is to protect the consumers' personal information and decrease credit card fraud cases to minimize financial risks. Due to the risk factors, customers often opt-out from paying via their credit cards [35]. Many researchers concluded that security is a significant factor influencing customer satisfaction in online

shopping [38]–[40]. A study found that behavioral intention to use digital wallets is strongly and vitally associated with perceived utility, perceived ease of use, and privacy and protection [21]. According to Peikari [24], security statements and technical protection significantly impact customer loyalty in the e-commerce industry but did not find a substantial effect on privacy.

Nevertheless, Barry and Jan [22] indicated that privacy and security positively correlate with behavior intention. Consumers may feel vulnerable to digital wallet transactions due to a lack of privacy and security. Nizam et al. [8] conducted empirical research to monitor digital wallets implementation in Malaysia. According to the study's findings, the dependent variable (customer purchasing decision using the digital wallet) is positively linked to all independent variables (convenience, security, and cost-saving).

Furthermore, convenience has the maximum significant positive correlation of 0.624, whereas security showed the coefficient correlation of 0.4999 with customer buying decisions using digital wallets. Based on the analysis, it was concluded that security has a more significant positive association with digital wallets for customer purchase behavior. In addition, Razif et al. [23] conducted empirical research among Malaysian young adults between 18 and 30 years old. The study showed that several factors have a significant relationship with the acceptance of the digital wallet platform. The factors listed were behavioral intention, perceived privacy risk, perceived usefulness, trust, perceived overall risk, and perceived performance risk.

Another empirical study found that trust, security, and privacy are the main factors affecting adopting a digital wallet [25]. A survey conducted by Subaramaniam et al. [13] demonstrated that security risk problems limit the prospect of using the digital wallet in Malaysia. According to Li et al. [26], cloud computing, security, e-learning, and quality of service are four significant factors that affect customer satisfaction in e-banking services. In addition, another research suggested that trust and privacy have positively impacted behavioral intention to use mobile banking services [27]. Likewise, Putra and Sfenrianto [28] demonstrated that a good payment system's security factor and speed influenced customer satisfaction in the digital payment method. Oliveira et al. [29] suggested that the digital payment system's security and performance directly affect customer loyalty. The research also indicated that security has a significant impact on mobile banking adoption. Customers expect banks to improve their security mechanisms by providing transaction security and privacy protection, particularly over wireless networks [9]. The level of security provided by a third-party online payment provider influences customer satisfaction [41]. In addition, Tang et al. [42] found that service quality, perceived risk, perceived security, perceived simplicity of use, social influence, and compatibility all substantially impact on consumers' intention to utilize digital payment.

A variety of factors influence customer intent to use e-wallets, including consumer perceptions of privacy, security.

Thus, the desire to use e-wallets is determined by the concern for transaction security and the protection of personal information given by users [43]. Furthermore, Qatawneh et al. [30] suggested that security and privacy statistically significantly impact the adoption of electronic payment system methods.

Previous researchers did not consider relevant security factors when assessing customer behavior but studied security as a general factor as shown in Appendix A. Besides, studies have shown that the security and privacy aspect of digital wallet systems when conducting transactions using digital wallet platforms is a primary concern for customers. The list of non-bank digital wallet issuers and the banks providing digital wallet services are shown in Appendix B and Appendix C, respectively.

The digital wallet services were developed for fast payment, school fee payments, handling expenses in fuel stations, global online shopping, NFC (Near Field Communication)-based transportation's payment, money transfer, bill payments, and others. Appendix B and Appendix C show 53 active digital wallet services, with 48 of them provided by Malaysian private and government fintech companies. The remaining five digital wallets are provided by international and local banks in Malaysia [44].

III. FRAMEWORK AND HYPOTHESES

Previous studies found that security substantially impacts consumer satisfaction in the digital wallet, but the studies examined security as an overall component. None of the research conducted identified the specific security factors in Malaysia's digital wallet. Thus, the current research proposed a six-factor security framework encompassing transaction speed, authentication, encryption mechanisms, software performance, privacy details, and information provided. Each factor was considered as a variable in this research. Fig. 1 represents the proposed framework of this study.

A. Transaction Speed

Transaction speed often refers to the rate at which data transfer happens from one record to another. The transfer speed is considered to be high if any transaction cannot be completed under a limited time period. An example of a real-life situation where transaction speed can be considered is the waiting time after the consumers have successfully paid for their online orders. The transaction speed of the payment application is a factor that may increase consumers' concerns [45]. The transaction speed is a characteristic that influences the development and satisfaction with any banking digital wallet technology. Numerous previous research found that transaction speed is a critical factor affecting consumer satisfaction with digital wallets [46]–[51]. Therefore, the following hypothesis is proposed:

- Hypothesis 1 (H1): There is a positive relationship between transaction speed and consumer satisfaction.

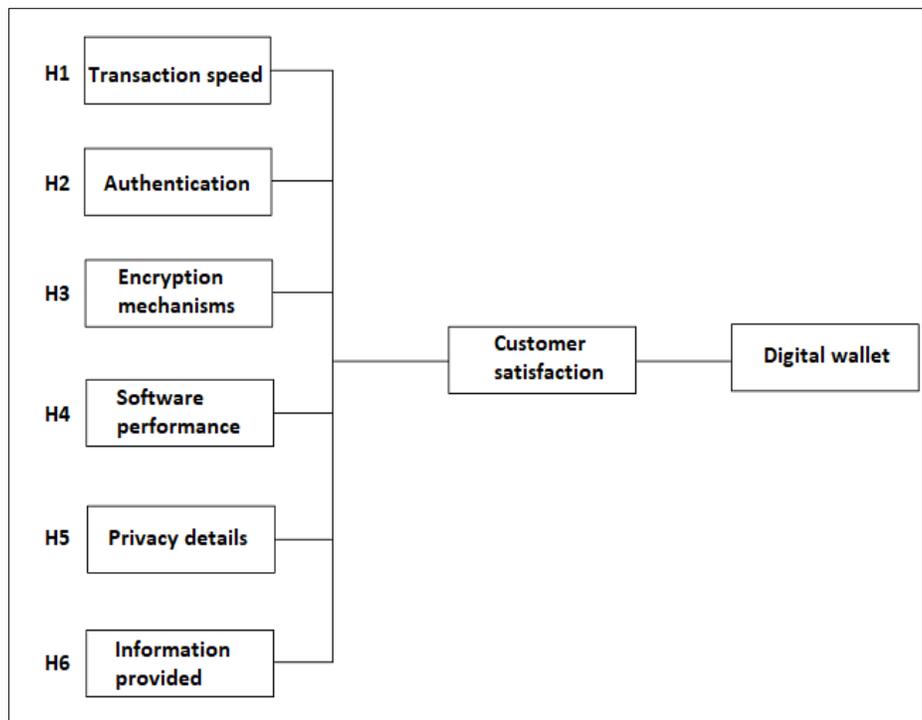


Fig. 1. The Proposed Framework for this Research.

B. Authentication

Authentication is a term that refers to the process of verifying a user's identification to guarantee that the activity being performed is being conducted by a trustworthy and real individual. It acts as a barrier to decrease the possibilities of identity theft. An exemplary situation of this would be the OTP code verification that consumers are required to do in order to complete their payment transactions. Authentication significantly impacts on consumer experience, which impacts their digital wallet adoption decision [52]–[54]. Because confidence is a significant influencing factor, digital wallet companies must guarantee that relevant aspects such as authentication are adequately regulated to build customer confidence [55]. Therefore, the following hypothesis is proposed:

- Hypothesis 2 (H2): There is a positive relationship between authentication and consumer satisfaction.

C. Encryption Mechanisms

Encryption mechanisms are often specific and unique steps and procedures done in turning order to encrypt data and ensure no third party or hackers can get the crucial information by encrypting the data into a gibberish form, which can only be decrypted using a unique key or mechanism corresponding to the encryption mechanism. Encryption mechanisms prevent hackers from breaking into a financial institution's server system. As a result, encryption mechanisms increase consumer confidence in conducting electronic payments [53], [56]. Therefore, the following hypothesis is proposed:

- Hypothesis 3 (H3): There is a positive relationship between encryption mechanisms and consumer satisfaction.

D. Software Performance

Software performance directs towards the overall performance of the software being used by consumers. In this case, performance acts as an indicator of how effectively the components and functions of the software meet their requirements. One of the most significant variables that directly influences acceptance intention for digital wallet is considered to be performance expectation [23], [57]–[59]. Incorrect functional usage situations and bugs in software may raise consumer worries. Therefore, the following hypothesis is proposed:

- Hypothesis 4 (H4): There is a positive relationship between software performance and consumer satisfaction.

E. Privacy Details

The information obtained from customers by digital services is referred to as privacy details in this context. Private information for registration purposes and authentication mechanisms are frequently included. Various previous studies have shown that customer satisfaction with digital wallets is significantly influenced by their ability to maintain their privacy [25], [60]–[65]. Therefore, the following hypothesis is proposed:

- Hypothesis 5 (H5): There is a positive relationship between privacy details and consumer satisfaction.

F. Information Provided

Because of the information provided by digital wallet services, customers of digital wallets may learn more about security. Customers may feel more confident about the security of the digital wallet system if they are informed of the security

procedures. Similarly, if users of digital wallets are unaware of security procedures, they may not feel secure [66]. Digital payments service knowledge has an important, positive, and simultaneous impact on the customer's ongoing desire to use digital wallet services [67]. The security information given by digital wallet services may thus assist customers in learning more about security and increasing their confidence in the system. Therefore, the following hypothesis is proposed:

- Hypothesis 6 (H6): There is a positive relationship between information provided and consumer satisfaction.

IV. METHODOLOGY

This research adopted an empirical research method. A quantitative online survey was distributed to the students and professors at the National University of Malaysia who fulfilled the criteria of used any digital wallet platform, including mobile applications and web-based systems, at least once. Overall, 300 responses were received. The survey was divided into two sections. The respondents were requested to fill in their demographic data in the first section, whereas the respondents responded to the questionnaire on customer satisfaction in the second section. In the personal information section, respondents provided demographic data, including gender, age, occupation, and the frequency of digital wallet transactions.

A five-point Likert scale was used in the second section for respondents to specify their level of agreement on a statement. The degree of agreement is used for the study's assessment process. There were 18 questions in the second section divided into six sections, respectively. Three questions were asked for every element proposed in the framework. Factor analysis, reliability analysis, and multiple regression analysis were conducted with the recollected questionnaires. IBM's Statistical Package for Social Science (SPSS) version 22.0 software was used to assess the statistical significance. Table I shows the respondents' characteristics.

As shown in Table I, the majority of respondents use digital wallet platforms frequently, and it is famous among youngsters.

TABLE I. CHARACTERISTICS OF RESPONDENTS

Indicator	Total cell	Accuracy (%)
Gender	Male	42%
	Female	58%
Age	18-25	61.80%
	26-35	21.80%
	36-45	9.10%
	46-More	7.30%
Occupation	Student	89.90%
	Professor	10.1%
Frequency of digital wallet transactions	More than 3 times a week	30.90%
	2-3 times a week	46.40%
	Once a week	22.70%
	Less than once a week	0%

V. RESULTS

A. Factor Analysis

Factor analysis is commonly employed in multivariate data analysis to evaluate the underlying dimensions [68]–[70]. It is a data rebate technique that transforms many variables into few variables. The highly influential variables were removed from the dataset during the factor analysis. The dataset was replaced with less influential variables after extracting the highly influential variables. The Kaiser-Meyer-Olkin (KMO) sampling adequacy test and Bartlett's test of sphericity were conducted to examine the dataset's construction validity. Table II demonstrates the KMO and Bartlett's test.

According to Table II, the average value of KMO is $.876 > 0.7$, whereas the Bartlett test's significance level of sphericity is 0.00, demonstrating that the collected data is normally distributed. The variables explained 87.6% of the variance from the total variance. Appendix D represents factor analysis. Each question posed for the six variables was used as a sub-variable to analyze the factors. Each factor was assigned with an abbreviation. TS indicated transaction speed, AT stands for authentication, EM denotes encryption mechanism, SP represents software performance, PD signifies privacy details, IP implies information provided, and CS designates customer satisfaction. Appendix D demonstrates that factor loading variables are more significant than 0.6. The minimum factor loading value is 0.656, and the maximum factor loading value is 0.965.

B. Reliability coefficient

The reliability test was performed using the SPSS software to test the dataset's internal consistency and obtain the Cronbach alpha coefficients. The values of the Cronbach alpha coefficients are illustrated in Table III.

As shown in Table III, the Cronbach alpha coefficients' values for all the variables measured are above 0.700. Cronbach alpha value of 0.700 or higher denotes an internally consistent dataset [71]–[73]. Therefore, the dataset received from the questionnaire survey satisfies the rule of thumb of validity.

TABLE II. KMO AND BARLETT'S TEST

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.876
Bartlett's Test of Sphericity	Sig.	.000

TABLE III. RELIABILITY COEFFICIENT

Variables	Number of Items	Reliability Coefficient (Cronbach Alpha)
Transaction speed	3	.912
Authentication	3	.851
Encryption mechanisms	3	.828
Software performance	3	.942
Privacy details	3	.935
Information provided	3	.902
Customer satisfaction	3	.917

C. Regression Analysis

Regression analysis is a practical way to analyze the variables and their connection [74]. The regression analysis is performed to identify the association between the dependent variable (customer satisfaction) and separate independent variables. The dataset’s model summary is shown in Table IV.

As shown in Table IV, the R² value is .723. The value represents a positive linear connection between customer satisfaction and other factors (independent variables) during the analysis. Table V shows the analysis of variance with a significant value less than 0.05.

Therefore, it is found that the independent variables influence the dependent variable. The descriptions of the coefficients are exhibited in Table VI. The variable with the highest β-Value in Table VI is relatively most important independent variable.

The negative value of constant exhibited in Table VI defines that when transaction speed, authentication, encryption mechanisms, software performance, privacy details, and information provided values are 0, the predicted value of customer satisfaction will be less than 0. The regression coefficient calculates a unit change in the dependent variable when the β-value represents independent variable change. Based on the β-value shown in Table VI, the extent of the independent variable effect on the dependent variable can be identified. A high β-value corresponds to high effects. Table VII shows the results for collinearity when multi-regression is applied. A Variance Inflation Factor (VIF) value greater than ten and a tolerance smaller than 0.2 implies a possible concern.

TABLE IV. MODEL SUMMARY

Model	R	R ²	Adjusted R ²	Std. Error of the Estimation
1	.850	.723	.717	.41674

TABLE V. ANALYSIS OF VARIANCE

Model	Sum of Squares	Df	Mean Square	F	Sig.
Regression	132.793	6	22.132	127.434	.000
Residual	50.887	294	.174		
Total	183.680	300			

TABLE VI. ANALYSIS OF COEFFICIENTS

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	β-Value
Constant	-0.770	0.181	
Transaction speed	0.176	0.044	0.173
Authentication	0.247	0.049	0.213
Encryption mechanisms	0.112	0.039	0.111
Software performance	0.215	0.045	0.189
Privacy details	0.128	0.034	0.139
Information provided	0.311	0.048	0.285

TABLE VII. COLLINEARITY STATISTICS

Model	T	Sig.	Collinearity Statistics	
			Tolerance	VIF
Constant	-4.248	.000		
Transaction speed	4.024	.000	.513	1.948
Authentication	5.040	.000	.530	1.887
Encryption mechanisms	2.865	.004	.630	1.587
Software performance	4.743	.000	.593	1.687
Privacy details	3.794	.000	.701	1.427
Information provided	6.529	.000	.498	2.008

According to Table VII, The VIF values are below 10, whereas the tolerance values are above 0.2 for all the independent variables. Hence, multi-regression is appropriate for the model, and there is no collinearity problem. The value of statistical significance (p-value) less than 0.05 indicates a statistically relevant correlation between the dependent and independent variables. Table VII indicates that the level of statistical significance for the independent variables is below 0.05. Table VIII shows the relationship between each variable. The Pearson correlation values indicate that the variables have a strong and moderate association with one another.

TABLE VIII. PEARSON CORRELATION

Pearson Correlation	TP ^a	Aut ^b	EM ^c	SP ^d	PD ^e	IP ^f	CS ^g
TP ^a	1						
Aut ^b	.563	1					
EM ^c	.460	.410	1				
SP ^d	.524	.508	.484	1			
PD ^e	.459	.407	.417	.431	1		
IP ^f	.591	.606	.509	.469	.366	1	
CS ^g	.675	.681	.572	.635	.537	.712	1

^a. TP = Transaction Speed

^b. Au = Authentication

^c. EM = Encryption mechanism

^d. SP = Software performance

^e. PD= Privacy details

^f. IP = Information provided

^g. CS = Customer satisfaction

D. Hypothesis Testing and Discussion

Table IX demonstrates the hypotheses testing. The test was conducted based on the data collected.

The hypotheses testing reveals that the p-value for the relationship between transaction speed and customer satisfaction is equal to 0.000, lesser than 0.05. Hence, H1 is supported. Therefore, it can be asserted that transaction speed has a significant positive impact on customer satisfaction. The findings suggest that students and academicians are more likely to use digital wallet systems if the transaction speed is high. According to the survey results, faster transaction speed will mitigate the security fears among digital wallet users. Users believe that fast online money transaction boosts the digital e-wallet platforms’ security.

TABLE IX. HYPOTHESIS TESTING

Hypothesis	Factor	β -Value	P-Value	Result
H1	Transaction speed	0.173	0.000	Supported
H2	Authentication	0.213	0.000	Supported
H3	Encryption mechanisms	0.111	0.004	Supported
H4	Software performance	0.189	0.000	Supported
H5	Privacy details	0.139	0.000	Supported
H6	Information provided	0.285	0.000	Supported

In addition, the p-value for the relationship between authentication and digital wallet customer satisfaction is less than 0.05 at 0.000. Thus, the H2 is also supported, denoting that authentication has a significant positive impact on customer satisfaction. This finding shows that user's digital wallet account authentication process influences the digital wallet system's consumer satisfaction. Digital wallet users believe that user authentication keeps scammers at bay and enhances digital wallet security.

Furthermore, the p-value for the relationship between encryption mechanisms and customer satisfaction is 0.004, which is less than 0.05. Therefore, H3 is supported. Thus, encryption mechanisms are found to exert a significant positive impact on customer satisfaction. The survey participants are concerned about accepting or denying digital wallet services with the encryption mechanisms. Similarly, the participants believe that a strong encryption mechanism will avoid abuse or hacking user information while using a digital wallet.

The p-value of software performance is 0.000, which is less than 0.05. Thus, the H4 is supported, indicating that software performance significantly impacts customer satisfaction. The findings confirm that digital wallet users are aware of software performance when using digital wallet platforms. Moreover, digital wallet users suspect that software with vulnerabilities increases the risk of digital wallet fraud.

Additionally, the p-value for the relationship between privacy details and customer satisfaction is 0.000, less than 0.05. Hence, H5 is also supported, concluding that privacy details significantly impact customer satisfaction. It suggests that private data collected from digital wallet users concern them. The digital wallet users opined that security vulnerabilities could be triggered by information collected through digital wallet platforms.

According to Table IX, H6 is also supported because the information provided has a p-value of 0.000, less than 0.05. Hence, the information provided has a significant positive impact on customer satisfaction. The finding shows that information provided by the digital wallet systems allows users of digital wallets to learn more about security. Providing additional security information increases the online payment systems' credibility. Furthermore, consumers will feel reassured about the digital wallet system's security when they are aware of software performance. Therefore, the study concluded that the proposed security factors significantly influence digital wallet consumer satisfaction based on the hypotheses tested.

From Table IX, among the studied factors, the information provided has the highest β -value (.285), indicating that the information provided to digital wallet users most significantly influence consumer satisfaction, followed by authentication (.213), software performance (.189), transaction speed (.173), privacy details (.139), and finally, encryption mechanisms (.111). From the analysis, the six influencing factors on customer satisfaction are arranged in ascending order according to their significant influences: Information Provided > Authentication > Software Performance > Transaction Speed > Privacy Details > Encryption Mechanisms.

VI. LIMITATION OF THE STUDY

Security was only covered as a general factor in previous studies. This study is the first research undertaken to identify the specific security factors for the digital wallet in Malaysia. However, this research was carried out based on responses from students and professors from the National University of Malaysia who may have better security factors awareness. Further research is required in this area of study. Security variables are complicated theoretical subjects for comprehension and research. An in-depth study of various populations is required to assess the factors suggested in this research. In addition, prospective researchers should also consider different security variables.

VII. CONCLUSION

This research has proposed a six-factor security framework that influences consumer satisfaction in Malaysia's digital wallet. Conclusively, all the proposed factors in the research have a significant positive influence on consumer satisfaction. Based on the analysis, the information provided most significantly influences customer satisfaction in digital wallets, followed by authentication, software performance, transaction speed, privacy details, and encryption mechanisms. Hence, improvised information security management principles are essential for the advancement of the digital wallet industry.

Digital wallets have gained popularity in recent times for providing cashless and comfortable daily payments or transactions. Although digital wallets deal with payments or transactions, research on considering and deducing security factors when developing digital wallet payment systems is limited. The progress of the digital wallet industry may impede without a thorough understanding of security factors. This research contributes to understanding the specific security factors necessary for financial technology companies. This study identifies new security factors that influence consumer satisfaction in digital wallet payment methods. The factors have not been analyzed in previous research. Therefore, this study contributes critically to the theoretical literature in digital wallet payments.

Consumer satisfaction is a crucial factor in the future for the booming digital wallet industry. Digital wallet security must be advanced to deal with emerging hackers and frauds. The outcome of this research can assist digital wallet providers in reinforcing the security of the system and focusing on crucial security factors to enhance customer satisfaction towards digital wallets. Moreover, this study can assist future

researchers planning to study this field by considering the variables proposed in this study.

ACKNOWLEDGMENT

This work was supported by the Ministry of Higher Education Malaysia (FRGS/1/2018/ICT01/UKM/02/5) and Universiti Kebangsaan Malaysia (GUP-2020-060).

REFERENCES

- [1] S. L. Miruna, "A Study on Customer Satisfaction Towards E - wallets in Tirunelveli City," *Ijciras*, vol. 2, no. 1, pp. 3–6, 2019.
- [2] M. Yang, A. Al Mamun, M. Mohiuddin, N. C. Nawi, and N. R. Zainol, "Cashless transactions: A study on intention and adoption of e-wallets," *Sustain.*, vol. 13, no. 2, pp. 1–18, 2021, doi: 10.3390/su13020831.
- [3] H. H. Bin Kadar, S. S. B. Sameon, M. B. M. Din, and P. 'Amirah B. A. Rafee, "Malaysia Towards Cashless Society," *Lect. Notes Electr. Eng.*, vol. 565, pp. 34–42, 2019, doi: 10.1007/978-3-030-20717-5_5.
- [4] Bernama, "Malaysia moving towards cashless society," *Free Malaysia Today*. 2017; Available from: <https://www.freemalaysiatoday.com/category/nation/2017/12/08/malaysia-a-moving-towards-a-cashless-society-say-bank-negara/> (accessed on 27 October 2021).
- [5] S. M. H. Mahmud, M. A. Kabir, O. A. M. Salem, and K. N. G. Fernand, "The comparative analysis of online shopping information platform's security based on customer satisfaction," *Proc. 2016 5th Int. Conf. Comput. Sci. Netw. Technol. ICCSNT 2016*, no. 2012, pp. 157–161, 2017, doi: 10.1109/ICCSNT.2016.8070139.
- [6] D. H. Shin, "Towards an understanding of the consumer acceptance of mobile wallet," *Comput. Human Behav.*, vol. 25, no. 6, pp. 1343–1354, 2009, doi: 10.1016/j.chb.2009.06.001.
- [7] M. J. López-Miguens and E. G. Vázquez, "An integral model of e-loyalty from the consumer's perspective," *Comput. Human Behav.*, vol. 72, pp. 397–411, 2017, doi: 10.1016/j.chb.2017.02.003.
- [8] F. Nizam, H. J. Hwang, and N. Valaei, "Measuring the effectiveness of E-wallet in Malaysia," vol. 786. Springer International Publishing, 2019.
- [9] S. Singh and R. K. Srivastava, "Predicting the intention to use mobile banking in India," *Int. J. Bank Mark.*, vol. 36, no. 2, pp. 357–378, 2018, doi: 10.1108/IJBM-12-2016-0186.
- [10] M. A. Hassan, Z. Shukur, M. K. Hasan, and A. S. Al-Khaleefa, "A Review on Electronic Payments Security," *Symmetry (Basel)*, vol. 12, no. 8, p. 1344, Aug. 2020, doi: 10.3390/sym12081344.
- [11] K. L. Y. Ming, M. Jais, C. C. Wen, and N. S. Zaidi, "Factor Affecting Adoption of E-Wallet in Sarawak," *Int. J. Acad. Res. Accounting, Financ. Manag. Sci.*, vol. 10, no. 2, pp. 244–256, 2020, doi: 10.6007/IJARAFMS/v10-i2/7446.
- [12] S. Luo, Y. Hu, and Y. Zhou, "Factors attracting Chinese Generation Y in the smartphone application marketplace," *Front. Comput. Sci.*, vol. 11, no. 2, pp. 290–306, Apr. 2017, doi: 10.1007/s11704-016-5022-8.
- [13] K. Subaramaniam, R. Kolandaisamy, A. Bin Jalil, and I. Kolandaisamy, "The impact of E-Wallets for current generation," *J. Adv. Res. Dyn. Control Syst.*, vol. 12, no. 1 Special Issue, pp. 751–759, 2020, doi: 10.5373/JARDCS/V12SP1/20201126.
- [14] N. Kumari and J. Khanna, "Cashless Payment: A Behaviourial Change To Economic Growth," *Qual. Quant. Res. Rev.*, vol. 2, no. 2, 2017, doi: 10.18535/ijre/v5i07.03.
- [15] Digital News Asia, "Almost 50% of Malaysian consumers distrust of online payment security," *Digit. News Asia 2018*; Available from: https://www.digitalnewsasia.com/digital-economy/almost-50-malaysian-consumers-distrust-online-payment-security?_cf_chl_jschl_tk__=pmd_1f36948d55bb8f989a3598e243b0478e129e9d4c-1627744177-0-gqNtZGzNAjjcnBsZQi6 (accessed on 31 July 2021).
- [16] M. Wolfenbarger and M. C. Gilly, "eTailQ: Dimensionalizing, measuring and predicting e-tail quality," *J. Retail.*, vol. 79, no. 3, pp. 183–198, 2003, doi: 10.1016/S0022-4359(03)00034-4.
- [17] M. Al-Okaily, A. Lutfi, A. Alsaad, A. Taamneh, and A. Alsyouf, "The Determinants of Digital Payment Systems' Acceptance under Cultural Orientation Differences: The Case of Uncertainty Avoidance," *Technol. Soc.*, vol. 63, no. September, 2020, doi: 10.1016/j.techsoc.2020.101367.
- [18] C. Ranganathan and S. Jha, "Examining online purchase intentions in B2C E-commerce: Testing an integrated model," *Inf. Resour. Manag. J.*, vol. 20, no. 4, pp. 48–64, Oct. 2007, doi: 10.4018/irmj.2007100104.
- [19] N. I. Ali, S. Samsuri, M. Sadry, I. A. Brohi, and A. Shah, "Online shopping satisfaction in Malaysia: A framework for security, trust and cybercrime," *Proc. - 6th Int. Conf. Inf. Commun. Technol. Muslim World, ICT4M 2016*, no. November, pp. 194–198, 2017, doi: 10.1109/ICT4M.2016.43.
- [20] Schaupp, L.C.; Bélanger, F. "A conjoint analysis of online consumer satisfaction," *J. of Electr. Commer. Res.* 2005; Vol. 6.
- [21] W. Karim, A. Haque, M. A. Ulfy, A. Hossain, and Z. Anis, "Factors Influencing the Use of E-wallet as a Payment Method among Malaysian Young Adults," *JIBM, J. Int. Bus. Manag.*, doi: 10.37227/jibm-2020-2-21.
- [22] M. Barry and M. T. Jan, "Factors Influencing the Use of M-Commerce: An Extended Technology Acceptance Model Perspective," *Int. J. Econ. Manag. Account.*, vol. 26, no. 1, pp. 157–183, 2018.
- [23] N. N. M. Razif, M. Misiran, H. Sapiri, "Perceived risk for acceptance of E-wallet platform in Malaysia among youth: Sem approach," *Manag. Res. J.*, vol. 9, pp. 1–24, 2020.
- [24] H. R. Peikari, "The influence of security statement, technical protection, and privacy on satisfaction and loyalty; a structural equation modeling," *Commun. Comput. Inf. Sci.*, vol. 92 CCIS, no. October, pp. 223–231, 2010, doi: 10.1007/978-3-642-15717-2_24.
- [25] E. D. Matemba and G. Li, "Consumers' willingness to adopt and use WeChat wallet: An empirical study in South Africa," *Technol. Soc.*, vol. 53, pp. 55–68, 2018, doi: 10.1016/j.techsoc.2017.12.001.
- [26] F. Li, H. Lu, M. Hou, K. Cui, and M. Darbandi, "Customer satisfaction with bank services: The role of cloud services, security, e-learning and service quality," *Technol. Soc.*, vol. 64, no. July 2020, p. 101487, 2021, doi: 10.1016/j.techsoc.2020.101487.
- [27] F. O. Bankole and O. O. Bankole, "The effects of cultural dimension on ICT innovation: Empirical analysis of mobile phone services," *Telemat. Informatics*, vol. 34, no. 2, pp. 490–505, 2017, doi: 10.1016/j.tele.2016.08.004.
- [28] H. R. Putra and Sfenrianto, "Analysis of customer satisfaction factors on e-commerce payment system methods in Indonesia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 4, pp. 471–480, 2020, doi: 10.14569/IJACSA.2020.0110463.
- [29] T. Oliveira, M. Thomas, G. Baptista, and F. Campos, "Mobile payment: Understanding the determinants of customer adoption and intention to recommend the technology," *Comput. Human Behav.*, vol. 61, no. 2016, pp. 404–414, 2016, doi: 10.1016/j.chb.2016.03.030.
- [30] A. M. Qatawneh, F. M. Aldhmour, and S. M. Alfugara, "The Adoption of Electronic Payment System (EPS) in Jordan: Case Study of Orange Telecommunication Company," *J. Bus. Manag.*, vol. 6, no. 22, pp. 2222–2847, 2015.
- [31] B. Liu, Y. Li, B. Zeng, and C. Lei, "An efficient trust negotiation strategy towards the resource-limited mobile commerce environment," *Front. Comput. Sci.*, vol. 10, no. 3, pp. 543–558, Jun. 2016, doi: 10.1007/s11704-015-4559-2.
- [32] K. B. Ooi and G. W. H. Tan, "Mobile technology acceptance model: An investigation using mobile users to explore smartphone credit card," *Expert Syst. Appl.*, Oct. 2016, vol. 59, pp. 33–46, doi: 10.1016/j.eswa.2016.04.015.
- [33] V. L. Johnson, A. Kiser, R. Washington, and R. Torres, "Limitations to the rapid adoption of M-payment services: Understanding the impact of privacy risk on M-Payment services," *Comput. Human Behav.*, vol. 79, pp. 111–122, 2018, doi: 10.1016/j.chb.2017.10.035.
- [34] K. Chan, C. Leong, B. Lim, and C. Yiong, "Sharing Economy through E-Wallet: Understanding the Determinants of User Intention in Malaysia," *J. Mark. Adv. Prat.*, vol. 2, no. 2, pp. 1–18, 2020.
- [35] M. S. M. Ariff, M. Sylvester, N. Zakuan, K. Ismail, and K. M. Ali, "Consumer perceived risk, attitude and online shopping behaviour: Empirical evidence from Malaysia," *IOP Conf. Ser. Mater. Sci. and Eng.*, June 2014, vol. 58, no. 1, p. 012007, doi: 10.1088/1757-899X/58/1/012007.

- [36] L. Yu, Z. Chen, P. Yao, and H. Liu, "A study on the factors influencing users' online knowledge paying-behavior based on the utaut model," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 5, pp. 1768–1790, 2021, doi: 10.3390/jtaer16050099.
- [37] M. F. Wei, Y. H. Luh, Y. H. Huang, and Y. C. Chang, "Young generation's mobile payment adoption behavior: Analysis based on an extended utaut model," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 4, pp. 1–20, 2021, doi: 10.3390/jtaer16040037.
- [38] Saylikhanov, S. "Factors Influencing Customer Satisfaction Towards Lazada Online Shopping in Malaysia," Available online: <http://studentrepo.iium.edu.my/handle/123456789/3184> (accessed on 31 July 2021).
- [39] S. S. Alam, M. H. Ali, N. A. Omar, and W. M. H. W. Hussain, "Customer satisfaction in online shopping in growing markets: An empirical study," *Int. J. Asian Bus. Inf. Manag.*, vol. 11, no. 1, pp. 78–91, 2020, doi: 10.4018/IJABIM.2020010105.
- [40] J. Kasuma, A. Kanyan, M. Khairoi, N. Sa'ait, and G. Panit, "Factors Influencing Customers Intention for Online Shopping," *Int. J. of Modern Trends in Bus. Res.* 2020, 3, 31-41.
- [41] Y. Choi and L. Sun, "Reuse intention of third-party online payments: A focus on the sustainable factors of alipay," *Sustain.*, vol. 8, no. 2, pp. 1–15, 2016, doi: 10.3390/su8020147.
- [42] Y. M. Tang, K. Y. Chau, L. Hong, Y. K. Ip, and W. Yan, "Financial innovation in digital payment with wechat towards electronic business success," *J. Theor. Appl. Electron. Commer. Res.*, vol. 16, no. 5, pp. 1844–1861, 2021, doi: 10.3390/jtaer16050103.
- [43] V. Soodan and A. Rana, "Modeling customers' intention to use e-wallet in a developing nation: Extending UTAUT2 with security, privacy and savings," *J. Electron. Commer. Organ.*, vol. 18, no. 1, pp. 89–114, 2020, doi: 10.4018/JECO.2020010105.
- [44] Non-bank E-money issuers. Bank Negara Malaysia. Available from: <https://www.bnm.gov.my/non-bank-e-money-issuers> (accessed on 31 July 2021).
- [45] R. Anjali and A. Suresh, "A study on customer satisfaction of bharat interface for money (BHIM)," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 6, pp. 266–273, 2019.
- [46] A. S. Yang, "Exploring Adoption difficulties in mobile banking services," *Can. J. Adm. Sci.*, vol. 26, no. 2, pp. 136–149, 2009, doi: 10.1002/cjas.102.
- [47] Ahmadinejad, B. "The impact of customer satisfaction on word of mouth marketing (Case study: Bamilo online store)," *SCIREA J. Manag.* 2019.
- [48] G. M. Ling, Y. S. Fern, L. K. Boon, and T. S. Huat, "Understanding Customer Satisfaction of Internet Banking: A Case Study In Malacca," *Procedia Econ. Financ.*, vol. 37, no. 16, pp. 80–85, 2016, doi: 10.1016/s2212-5671(16)30096-x.
- [49] A. M. Khalaf Ahmad and H. Ali Al-Zu'bi, "E-banking Functionality and Outcomes of Customer Satisfaction: An Empirical Investigation," *Int. J. Mark. Stud.*, vol. 3, no. 1, pp. 50–65, 2011, doi: 10.5539/ijms.v3n1p50.
- [50] M. Jannat and I. Ahmed, "Factors Influencing Customer Satisfaction of Mobile Banking Services: A Study on Second - Generation Banks," *Eur. J. Bus. Manag.*, vol. 7, no. 26, pp. 88–97, 2015.
- [51] N. Jahan, M. J. Ali, and A. Al Asheq, "Examining the key determinants of customer satisfaction internet banking services in Bangladesh," *Acad. Strateg. Manag. J.*, vol. 19, no. 1, pp. 1–6, 2020.
- [52] Cheah, J.S.; Isa, S.M.; Yang, S. "The Impact of Perceived Usefulness, Perceived Value, and Perceived Security on Mobile Payment App Loyalty through Satisfaction: User Interface as Moderator," *Proc. The 41th Nat. and Int. Conf. Glob. Goals, Loc. Act. Look. Back and Mov. For.* 2021, 1, 14.
- [53] N. N. Duy Phuong, L. T. Luan, V. Van Dong, and N. Le Nhat Khanh, "Examining customers' continuance intentions towards e-wallet usage: The emergence of mobile payment acceptance in Vietnam," *J. Asian Financ. Econ. Bus.*, vol. 7, no. 9, pp. 505–516, 2020, doi: 10.13106/JAFEB.2020.VOL7.NO9.505.
- [54] C. S. Weir, G. Douglas, M. Carruthers, and M. Jack, "User perceptions of security, convenience and usability for ebanking authentication tokens," *Comput. Secur.*, vol. 28, no. 1–2, pp. 47–62, 2009, doi: 10.1016/j.cose.2008.09.008.
- [55] Bhatt, V. "Factors affecting the consumer's adoption of E -wallets in India : An empirical study," *SAL. Inst. of Man.* 2020, 9 , 6.
- [56] S. Phophalia, G. Goswami, M. Prasad, M. Arora, and B. Graph, "A Study on Impact on Customer Satisfaction for E-Wallet Using Path Analysis model," *J. Bank. Insur. Law, no. Iccm*, 2018.
- [57] X. Luo, H. Li, J. Zhang, and J. P. Shim, "Examining multi-dimensional trust and multi-faceted risk in initial acceptance of emerging technologies: An empirical study of mobile banking services," *Decis. Support Syst.*, vol. 49, no. 2, pp. 222–234, 2010, doi: 10.1016/j.dss.2010.02.008.
- [58] G. Baptista and T. Oliveira, "Understanding mobile banking: The unified theory of acceptance and use of technology combined with cultural moderators," *Computers in Human Behavior*, vol. 50, pp. 418–430, 2015, doi: 10.1016/j.chb.2015.04.024.
- [59] T. Zhou, Y. Lu, and B. Wang, "Integrating TTF and UTAUT to explain mobile banking user adoption," *Comput. Human Behav.*, vol. 26, no. 4, pp. 760–767, 2010, doi: 10.1016/j.chb.2010.01.013.
- [60] H. Alnaaji and A. Qusef, "A Conceptual Framework for Representing Business Functions and Their Role in the Quality of E-Banking Services," 2021 Int. Conf. Inf. Technol. ICIT 2021 - Proc., pp. 627–633, Jul. 2021, doi: 10.1109/ICIT52682.2021.9491726.
- [61] M. F. Putri, B. Purwandari, and A. N. Hidayanto, "What do affect customers to use mobile payment continually? A systematic literature review," 2020 5th Int. Conf. Informatics Comput. ICIC 2020, Nov. 2020, doi: 10.1109/ICIC50835.2020.9288590.
- [62] C. Ayu, K. Larasati, and R. A. Salim, "Analysis of Factors Influencing Continuance Intention of E-wallet Use : A Case Study of LinkAja," *Irjaes*, vol. 6, no. 2, pp. 27–33, 2021.
- [63] C. Mombeuil, "An exploratory investigation of factors affecting and best predicting the renewed adoption of mobile wallets," *J. Retail. Consum. Serv.*, vol. 55, no. April, p. 102127, 2020, doi: 10.1016/j.jretconser.2020.102127.
- [64] M. Darmiasih and P. Y. Setiawan, "Continuance usage intention and its antecedents on using OVO e-wallet application in Denpasar," *Int. Res. J. Manag. IT Soc. Sci.*, vol. 8, no. 1, pp. 35–46, 2020, doi: 10.21744/irjmis.v8n1.1104.
- [65] P. Sarika and S. Vasantha, "Review on Influence of Trust on Mobile Wallet Adoption and its Effect on Users' Satisfaction," *Int. J. Manag.*, vol. 8, no. 1731, pp. 1731–1744, 2018.
- [66] Akhila Pai H. Study on consumer perception towards digital wallets. *Int. J. Res. Anal. Rev.* 2018, 5, 385–391.
- [67] Lim, S.H.; Kim, D.J.; Hur, Y.; Park, K. "An Empirical Study of the Impacts of Perceived Security and Knowledge on Continuous Intention to Use Mobile Fintech Payment Services," *Int. J. of Human Comput. Inter.* 2018, 35, 886–898, doi:10.1080/10447318.2018.1507132.
- [68] N. Ghazali and M. S. Nordin, "Measuring meaningful learning experience: Confirmatory factor analysis," *Int. J. Innov. Creat. Chang.*, vol. 9, no. 12, pp. 283–296, 2019.
- [69] S. McQuitty, "The Purposes of Multivariate Data Analysis Methods: an Applied Commentary," *J. African Bus.*, vol. 19, no. 1, pp. 124–142, 2018, doi: 10.1080/15228916.2017.1374816.
- [70] Bornmann, L.; Haunschild, R. Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000 prime data. *PLoS one* 2018, 13, e0197133, doi:10.1371/journal.pone.0197133.
- [71] Nunnally JC. "Psychometric theory 3E," Tata McGraw-hill education 1994.
- [72] Van Griethuijsen, R.A.L.F., van Eijck, M.W., Haste, H. et al. "Global Patterns in Students' Views of Science and Interest in Science," *Res Sci Educ.*, 2015, 45, 581–603, doi:10.1007/s11165-014-9438-6.
- [73] K. S. Taber, "The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education," *Res. Sci. Educ.*, vol. 48, no. 6, pp. 1273–1296, 2018, doi: 10.1007/s11165-016-9602-2.
- [74] Fang, S. Empirical Study of Influential Elements of University Students' E-satisfaction. *Int. J. of Bus. and Soc. Sci.* 2014, 5, 6.

APPENDIX A

Ref	Objective	Finding	Limitation
[8]	Identify the significant factors affecting the purchasing intention of customers using E-wallet in Malaysia.	Convenience, cost-savings, and security were identified to impact customer purchasing behavior using E-wallet.	Security factors were not considered as variables.
[9]	Analyze the variables affecting the customer's decision to use mobile banking in India.	Perceived ease of use, usability, social effects, security, and perceived cost affect the decision of customers to use mobile banking.	The influence of security variables was not addressed in this study.
[11]	Examine the variables that impact the acceptance of E-wallet services in Sarawak, Malaysia.	The acceptance of e-wallet services is influenced by perceived risk, perceived usefulness, and perceived ease.	Security variables were not considered to identify the perceived risk factors.
[13]	Assess the positive and negative effects of e-wallet on Malaysian users.	The limitations of using the digital wallet in Malaysia are technological challenges and security risks.	The analysis did not classify digital wallet security risks but instead considered security as a general aspect.
[17]	Explain and recognize the nature of adopting the digital payment system in Jordan on the framework of the UTAUT2 model.	Performance expectations, social effect, price value, security, and privacy were important digital payment system acceptance indicators.	Security factors were not considered as variables.
[18]	Identify the relative primary factors that impact online purchase intentions.	Privacy, security, and delivery have a significant positive impact on customer satisfaction when it comes to online purchase intention.	The study does not identify the specific security factors for digital wallet systems.
[19]	Proposed a framework of the security factors in online shopping.	This framework has five security factors: financial security, privacy, system security, cybercrime, trust, and customer satisfaction.	The study does not identify the specific security factors for digital wallet system.
[20]	An analysis to determine the attributes that impact customer satisfaction in online shopping.	Privacy, merchandising, convenience, trust, delivery, usability, product customization, product quality, and security are the important attributes to the consumer for online satisfaction.	Specific security factors were not considered as variables.
[21]	Identify the primary factors that contribute to the acceptance of electronic payment systems.	Compatibility, the perceived security of technology, performance expectations, creativity, and social impact have important beneficial and detrimental effects on the acceptance and recommendation of electronic payment systems.	The analysis did not consider security factors but rather addressed security as a general factor to assess consumer behavior.
[22]	Investigate the variables that have a significant impact on the adoption of Jordan's electronic payment systems.	Security and privacy have a statistically significant impact on the adoption of electronic payment system methods.	This research does not identify particular security factors to assess the behavior of consumers to prevent ambiguity.
[23]	Examine the driving factors over the use of e-wallet as a payment method by Malaysian young adults.	Perceived usefulness, perceived ease of use and privacy and security have a positive and vital association with the e-wallet behavioral intention.	Security was considered as a general factor; specific security variables were not studied.
[24]	Analyze the factors influencing m-commerce use in Malaysia.	Perceived usefulness, perceived satisfaction, security, and privacy have a significant positive impact on the behavioral intention of m-commerce use in Malaysia.	The study does not consider specific security factors as variables. Security was considered a general factor.
[25]	Analyze the perceived risk that represents the ambiguity of adverse effects over e-wallets in Malaysia on consumer emotions.	The perceived risk of privacy, perceived usefulness, trust, perceived general risk, and perceived risk of performance are directly linked to the acceptance of the e-wallet platform.	The study does not specify the specific security factors for the digital wallet system.
[26]	Investigate the influence of security statements, technical protection, trust, and privacy on customer satisfaction, in the world of e-commerce.	Security statements and technical protection significantly impact customer loyalty in the e-commerce industry, and the study found no substantial effect on privacy.	The study considered security statements as a variable and did not study specific security factors.
[27]	Predict the extent of adoption of the People-to-People (P2P) services of the WeChat wallet in South Africa.	Trust, security, and privacy impact the decisions of South Africans to accept the WeChat wallet.	The research was based on WeChat wallet mainly, did not include other e-wallet services, and did not analyse security factors.
[28]	Examine the variables impacting consumer satisfaction with e-banking systems.	Cloud computing, security, e-learning, and quality of service are factors that can improve customer loyalty with e-banking.	The study did not specify digital wallet security variables.
[29]	Examine the influence of socio-cultural factors on ICT innovation, emphasizing mobile banking services in South Africa.	Trust and privacy have a significant impact on the behavior intention of using mobile banking services.	The analysis does not identify the security factors that affect consumer behavior.
[30]	Identify the variables that affect consumer satisfaction on the systems of electronic payment in Indonesia.	Customer loyalty has a significant influence on service effectiveness, benefits provided, the security of transactions, speed, active usage, benefits received, and convenience of transactions.	The analysis did not classify digital wallet security risks but rather considered the security of transactions as a general aspect.
[35]	Identify the variables that impact consumer perceived risk and their attitude toward online shopping in Malaysia.	The attitude of online shoppers is adversely influenced by product risk, financial and non-delivery risks. Convenience risk was shown to have a positive impact on the mindset of the customer.	The study does not identify the security factors that impact consumer behavior.

[36]	Identify the variables that impact users' paying behavior.	Content quality, peer influence, KOL influence, perceived interaction, effort expectation, and perceived trust all have a substantial impact on users' paying behavior.	The findings of the research do not identify which security variables influence customer behavior.
[37]	Examine the factors that influence the satisfaction of the younger generation with digital payment systems.	Perceived risks are shown to have a significant impact on the behavior intention of using digital payment services.	The research did not identify any factors relating to the security of digital wallets.
[42]	Analyze consumers' intention to utilize digital payment.	Service quality, perceived risk, perceived security, perceived simplicity of use, social influence, and compatibility all have a substantial impact on consumers' intention to utilize digital payment.	The findings of the study do not reveal which security variables influence consumers' purchase decision. In this case, security was taken into consideration in a wide sense.

APPENDIX B

No.	Issuers (Non-Banks)	Wallet name	No.	Issuers (Non-Banks)	Wallet name
1	AEON	AEON Member, Plus Card	25	Mobile Money International	Money Pin
2	Alipay Malaysia	Lazada Wallet	26	MobilityOne	eM-onei
3	Axiata Digital eCode	Boost	27	MOL AccessPortal	Razer Gold
4	Bandar Utama City Centre	IPAY	28	MRuncit Commerce	Mcash
5	Bayo Pay (M)	Construx	29	MyEG Alternative	iPayEasy
6	BigPay Malaysia	BigPay	30	TNG Digital Remittance	NAPP (Numoni App)
7	BLoyalty	B Infinite Pay	31	PayPal Pte. Ltd	PayPal
8	Chevron Malaysia Limited	Caltex StarCard Debit	32	Petron Fuel International	Petron Prepaid Fleet Card
9	DIV Services	Whalet	33	Presto Pay	Presto Pay
10	Fass Payment Solutions	Fasspay	34	qBayar	qBayar
11	Finexus Cards	Visa / Master Prepaid Card	35	Raffcomm	e-Info
12	Fullrich Malaysia	TaPay	36	Razer Pay Wallet (M)	Razer Pay
13	Gkash	Gkash eWallet	37	Serba Dinamik IT Solutions	Qwikpay
14	Google Payment Malaysia	Google Play Gift Card	38	Setel Ventures	Setel App
15	GoPay	GoPay	39	ScanPay	MyScanPay
16	GPay Network (M)	GrabPay	40	ShopeePay Malaysia	ShopeePay
17	Instapay Technologies	Instapay e-Wallet	41	SiliconNet Technologies	Sarawak Pay
18	iPay88 (M)	iPay88 e-Wallet	42	SMJ Teratai	eWANG
19	I-Serve Payment Gateway	Zapp	43	Touch 'n Go	Touch 'n Go, Prepaidcard
20	JuruQuest Consulting	QBpay e-wallet	44	TNG Digital	Touch 'n Go eWallet
21	KiplePay	kiplePay	45	U Mobile Services	GoPayz
22	ManagePay Services	Mpay	46	Wavpay Systems	Wavpay
23	Maxis Broadband	Prepaid Airtime	47	WeChat Pay Malaysia	WeChat Pay
24	Merchantrade Asia	Valyou Wallet	48	XOX Com	XOX eWallet

APPENDIX C

No.	Banks	Products
1	AmBank (M) Berhad	Prepaid Card (MasterCard)
2	Bank of China (M) Berhad	Prepaid Card (China Union Pay)
3	CIMB Bank Berhad	Prepaid Card (MasterCard) CIMB Pay
4	Malayan Banking Berhad	QR Pay
5	RHB Bank Berhad	Prepaid Card (Visa)

APPENDIX D

Variables	ID	Measurements Items	Values
Transaction speed	TS1	Slow online money transaction speed can increase the chances of becoming a fraud victim while making payments using a digital e-wallet.	.816
	TS2	Fast online money transaction speed improves the security of the digital e-wallet platform.	.847
	TS3	A faster online money transaction speed gives hackers less time to commit fraud.	.965
Authentication	AT1	User authentication has a directly proportional relationship with digital e-wallet security.	.880
	AT2	User authentication helps in ensuring the genuine cardholder is in charge while completing transactions online.	.846
	AT3	User authentication acts as another form of measure to keep scammers away.	.768
Encryption mechanisms	EM1	A good encryption mechanism can prevent the user information from being misused or hacked.	.656
	EM2	An encryption mechanism acts as a barrier between the customer and third parties with malicious intent to steal the customer information.	.681
	EM3	Encrypted data would have no value when stolen by a hacker because the data is encrypted.	.816
Software performance	SP1	A software with bugs increases the chances of fraud in the digital wallet.	.939
	SP2	The higher and better a software's performance, the harder it is for a hacker to break in.	.877
	SP3	A software with a slower performance gives a bigger scope for hackers to find the defects in the system.	.916
Privacy details	PD1	Information taken from the user can cause security issues perceived risk.	.868
	PD2	User's information is vulnerable.	.877
	PD3	The more confidential information stored results in a higher user perceived risk.	.954
Information provided	IP1	Information provided by the digital wallet system can help the user to understand more about security.	.923
	IP2	Providing more information about security improves the transparency of an online payment system.	.834
	IP3	Users will feel more assured and at ease if they are provided with more security information.	.880
Customer satisfaction	CS1	Digital wallet services have accelerated my regular activities.	.860
	CS2	Compared to conventional techniques, the digital wallet is a time-saving system.	.872
	CS3	I expect that in the near future, I would be using digital wallet systems.	.842

Towards a Strategic IT GRC Framework for Healthcare Organizations

Fawaz Alharbi¹

Computer Science Department, Huraymila College of
Science and Humanities, Imam Mohammad Ibn Saud
Islamic University, Riyadh, Saudi Arabia

Mohammed Nour A. Sabra²

Clinical Review Department
King Fahad Medical City
Riyadh, Saudi Arabia

Nawaf Alharbe³

Computer Science Department
Applied College, Taibah University
Madina, Saudi Arabia

Abdulrahman A. Almajed⁴

Governance, Risk and Compliance Department
Ministry of Industry and Mineral Resource
Riyadh, Saudi Arabia

Abstract—The rapidly changing healthcare market requires healthcare institutions to adjust their operations to address regulatory, strategic, and other risks. Healthcare organizations use a wide range of IT systems producing large amounts of sensitive and confidential data. However, few tools are available to measure the data governance activities of healthcare institutions and align healthcare data management with legislation. The Governance, Risk, and Compliance (GRC) Model focused on integrating that ability to achieve organizational goals. The demand for corporate governance is crucial for protecting the healthcare system from risks. An adaptation of a modified version that includes strategy, processes, technology, people, as well as legal and business requirements was developed to analyze the factors affecting IT GRC implementation in healthcare organizations. Although about 48% of participants reported that their organizations implemented IT GRC programs, 16% stated that they are considering implementing IT GRC programs soon. In almost 71% of healthcare organizations, IT governance, risk management, and compliance are integrated. Among the factors influencing the implementation of IT GRC programs in Saudi healthcare organizations, legal context ranked as the most critical, followed by process, strategy, then technology, business, and finally, people contexts. This study shows that healthcare organizations must assess various factors for the effective implementation of IT GRC activities.

Keywords—Information technology; strategic; healthcare; governance; compliance

I. INTRODUCTION

Increasing economic uncertainty, evolving market trends, and expanding regulations are escalating health organizations' risk exposure [1]. Currently, changes in healthcare systems are disturbing operations, necessitating the implementation of effective risk management systems. Hence, maintaining competitiveness and managing risk in the healthcare environment requires new actions, plans, and strategies. These changes have been facilitated by updated government regulations, organizational structures, accountability measures, and the relationship between consumers and healthcare providers [2]. In the evolving healthcare market, institutions must modify their operations to address regulatory, strategic,

and other risks. However, the need to change these models, along with the implementation of new models and organizational structures, might increase industry risks while providing lucrative opportunities. As a result, healthcare organizations are learning how to use effective models to turn those risks into profits.

The healthcare domain contains many information systems that produce massive amounts of data. These data include patient data, disease research data, healthcare professional data, and other sensitive information that must be managed with the highest levels of confidentiality, integrity, and availability (CIA) [3]. Many healthcare organizations realize that the changes in healthcare systems are disrupting their usual practices, and they need effective risk management for data governance. They also recognize that maintaining competitiveness and managing risks in the healthcare environment require new actions, plans, and strategies. The modification in government regulations, organizational structures, and accountability measures as well as the relationship between consumers and healthcare providers [4], facilitate these changes. However, there is a lack of developed tools for measuring data governance activities and aligning the healthcare organization data management with legislation.

One of the models that have been introduced in the healthcare sector to deal with these issues is Governance, Risk, and Compliance (GRC). GRC focuses on integrating a range of capabilities that assist health organizations to act with integrity, maintain consistency, address uncertainty, and achieve organizational goals [5]. As the dependency on information systems increased, the need to apply the GRC concept to IT operations became apparent. IT GRC ensures that all IT systems have proper governance, risk management, and compliance management to support healthcare organizations [6]. Although many researchers identified the need for IT GRC in healthcare organizations, only a few addressed the need to develop a strategic framework for implementing IT GRC in healthcare settings [7].

The benefits of applying IT GRC in healthcare organizations include accountability, better management for

electronic health records, alignment with legal requirements. The purpose of this study is to support healthcare organizations in identifying IT GRC practices at a strategic level. This paper evaluates the main factors affecting IT GRC in the healthcare sector to develop a strategic model that addresses various perspectives.

II. BACKGROUND

A. Factors Led to the Emergence of IT GRC in the Healthcare Sector

The emergence of IT GRC as an approach for protecting healthcare organizations from excessive risk and removing growth barriers has been due to numerous factors. One such factor is the demand for corporate governance. Governance is a broad term that involves individuals who administer the operations, laws, processes, institutions, and policies which define the structure that manages and directs healthcare organizations [8]. Thus, governance affects how healthcare organizations address everything, such as daily operations and patient care strategies. If any organizational operation fails, the board of directors and executives are held accountable rather than the policies or organizational culture [9]. Therefore, an accountability issue arises whether external and internal constituencies trust that the healthcare providers are doing everything to protect the quality of care and mitigate risk. Recently, there has been increased media coverage of healthcare organizations that fail to ensure the safety and protection of sensitive patient information [4]. Due to increased media attention, data breach, and the increasingly complex healthcare regulatory environment, board members and executives are more thoroughly accustomed to how their healthcare organizations operate. This increased scrutiny has ensured that the individuals in the governance have timely and accurate information regarding their organization. These individuals have the capability of making decisions that ensure compliance, prevent unnecessary risk, and reduce the impacts and chances of regulatory penalties and patient litigation [10].

Another factor that has led to the emergence of IT GRC in the healthcare sector is the increased adoption of electronic health records (EHR) that has led to greater risk. There is a growth of data in the healthcare sector, such as patient information. This growth indicates that healthcare organizations need an effective structure that clearly illustrates business requirements, data governance, and technology processes and infrastructure to support a secure data management environment [11]. Nonetheless, some healthcare organizations face challenges, such as increasing regulatory standards and requirements, a lack of funds for security initiatives, and the growing need for data sharing among collaborators and partners. These challenges have made it difficult for organizations to protect and manage all these data centers. Numerous areas have increased the risk due to the adoption of EHR, such as the global and dynamic nature of electronic information, collaborative patient care, and the utilization of electronic patient portals [3]. Due to the complexity of safeguarding health information, there is a need to have a more holistic risk management approach, such as carrying out a risk assessment on the electronic environment.

Another contribution to the emergence of IT GRC in the healthcare sector is the growth of regulatory requirements. Healthcare organizations must comply with these regulatory requirements to minimize the impact and chances of regulatory penalties and patient litigation [12]. In the healthcare sector, compliance denotes the act of adherence to regulations, along with the capability of healthcare providers to demonstrate and sustain that. Healthcare organizations need to adhere to internal policies as well as externally imposed regulations and laws. In addition to regulatory requirements, there are a more informed public, more assertive regulators, and more serious non-compliance penalties, which indicate that the organization needs to focus on compliance [13].

The boundaries of the extended healthcare enterprise are disappearing due to the far-reaching and intricate web of relationships [14]. For instance, numerous departments and constituencies share patient information. Additionally, the use of advanced technologies such as VoIP services, mobile devices, social networking, virtualization, and cloud outsourcing has led to the disappearance of conventional boundaries of a single healthcare enterprise [15]. This cross-pollination of information and services has made it difficult to determine the beginning and end of one healthcare operation. It has also led to unwanted risks. It is challenging to manage the numerous healthcare sectors that affect patient information. Additionally, many divisions in the healthcare organizations, such as radiology, the ER, and hospital labs, have burdened healthcare providers with assessments of employment practices, workflow support, privacy, security, and health and safety. Therefore, healthcare organizations need to validate their extended enterprise members to meet social responsibility practices, comply with laws, and ensure that they operate in a manner that prevents unnecessary risks. This is because new technologies generate numerous opportunities and risks [16]. For instance, the utilization of mobile devices by healthcare providers might affect the delivery of services along with creating new expectations across numerous touchpoints. At every touchpoint, there is a likelihood of introducing risks relying on the capability of health organizations to secure the connection. Hence, the disappearing boundaries in healthcare enterprises that might lead to unwanted risk have contributed to the emergence of IT GRC.

B. IT GRC in Healthcare Settings

Researchers in [17] provided a clear GRC model based on the maturity model. The authors indicated that the GRC framework is most common among multinationals, insurance, banking, and listed corporations. However, the article acknowledged that there is a need for the GRC framework in healthcare settings. [18] indicated that in past years the GRC framework was not adopted in healthcare, although it has gained popularity among developed countries such as Germany and the U.S. The author discussed only the current state and significance of GRC in healthcare care without offering a clear IT GRC framework. [19] discussed past and future directions of information technology governance. The study did not provide a clear IT GRC framework that healthcare organizations can implement. [20] provided a clear guideline on how governance, risk, and compliance can align to ensure better decision-making. The study has indicated that, in past

years, organizations have failed to align governance, risk, and compliance. However, the study did not provide a complete GRC framework that healthcare organizations can adopt specifically for the IT GRC setting. Table I summarizes the related literature.

The literature agrees that there is a need to develop an IT GRC framework that healthcare organizations can use.

TABLE I. SUMMARY OF RELATED LITERATURE

Reference	Contribution	Limitation
[17]	GRC maturity model for hospitals	Not specific for IT GRC in healthcare organizations
[18]	Advantages of implementing GRC in healthcare care systems	No clear IT GRC framework for healthcare organizations
[19]	Past and future directions of information technology governance	No clear IT GRC framework for healthcare organizations
[20]	Clear guideline on how governance, risk, and compliance can be aligned for healthcare organizations	Partial GRC framework for healthcare organizations not specific for IT GRC

III. RESEARCH FRAMEWORK

This research aims to assess the factors affecting the implementation of IT GRC initiatives in healthcare organizations. This paper adopted a modified version of the frame of reference for integrated GRC developed in [2]. The original frame contained four components, strategy, processes, technology, and people. This research framework adds two new components, business and legal requirements. The research framework appears in Fig. 1. The description of the research framework components are as follows:

- **Strategy:** The alignment among healthcare strategy, IT strategy, and IT GRC activities is crucial for integrating IT GRC in healthcare organizations [21]. A study conducted in Swiss hospitals found that IT directors usually make all the decisions without any discussion with related departments in about 75% of the hospitals [1].
- **Process:** IT GRC involves many processes that span various GRC domains (i.e., governance, risk management, compliance) [22]. IT governance processes must control IT risk management and IT compliance management in the healthcare organization [23].
- **Technology:** Technology plays a vital role in the integration of IT GRC activities. However, applying IT GRC technical tools in the healthcare organization requires an understanding of the nature of the healthcare business [24].
- **People:** People play different roles in IT GRC activities, such as identifying risks and managing the systems [23]. Understanding the importance of IT GRC activities in healthcare organizations is one of the success factors for implementing any IT GRC initiative in the healthcare organization [24].

- **Business:** Implementing IT GRC requires investment in technical solutions and processes. Thus, successful implementation requires approval from both business and information technology leaders to make the implementation successful [1].
- **Legal:** Regulatory compliance is one of the major drivers for any GRC program [25]. In the healthcare domain, organizations have to comply with various regulations. While some of these regulations are healthcare-specific, others could be related to IT regulations or other general laws.

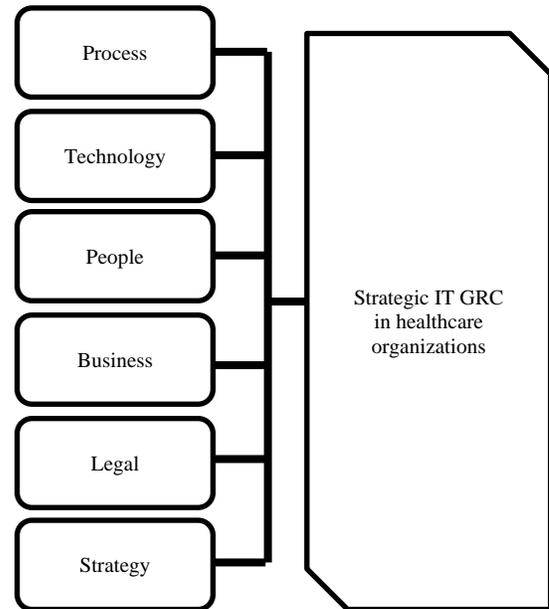


Fig. 1. Research Framework.

IV. RESEARCH METHODOLOGY

To assess the factors affecting the implementation of IT GRC initiatives in healthcare organizations, this study started with a literature review, followed by a quantitative survey. The questionnaire items fall into six categories (technology, process, people, strategy, legal requirements, and business) see Appendix 1.

Two experts (a professor and an assistant professor) reviewed the developed questionnaire to test the content validity. After making the required changes, the authors translated the questionnaire into Arabic. Then, the research was reviewed by the Research Ethics Committee at Shaqra University in Saudi Arabia to gain ethical approval.

The research team used Microsoft Forms as an online tool to distribute the questionnaire. The targeted population of this research is all staff members in Saudi healthcare organizations who are responsible for IT GRC-related activities or similar roles. The researchers used the snowball technique to reach all possible participants.

The questionnaire consists of five parts. The first part acts as a cover letter by providing information about the study and the researchers and taking the informed consent of the participants. The second part collects information about IT

GRC activities inside the healthcare organizations, such as the type of IT GRC program in the organization and who is in charge of the IT GRC program in the organization. The third part includes statements developed to assess the factors affecting the implementation of IT GRC initiatives in healthcare organizations. The fourth part collects the demographic information of the participants. The final part allows the respondents to add additional comments regarding the topic.

V. DATA ANALYSIS

After completing all the required processes for the questionnaire development, the questionnaire was distributed to the targeted audiences of Saudi healthcare organizations members. We received 122 responses from 19 Oct 2021 to 7 Dec 2021. Four of the respondents were not accepted due to incompleteness or because they worked in the wrong industry. The analysis was conducted using Jamovi software [26]. This study implies quantitative methods as the analysis methods for many reasons such as the ability to generalization, high level of objectivity.

A. Sample Characteristics

Table II shows the demographic information of the participants. The gender balance among the participants was almost equal (Male 52.5%, Female 47.5%). The majority of the participants work in healthcare organizations in the Central Province (66.1 %) and the Western Province (25.4%). This was expected since most of the main locations of the healthcare organizations in Saudi Arabia reside in those two provinces, which have the majority of the Saudi population [27]. Although 67% of the participants have five years or more of experience in the healthcare sector, only 21.1% have experience with more than five years in IT GRC related activities.

B. IT GRC Status in Healthcare Organizations

This section discusses the IT GRC status in Saudi healthcare organizations. While 48.3% of the participants reported that their organizations implemented the IT GRC program, only 16.1% stated that their organization plans to have an IT GRC program in the foreseeable future.

High percentages of participants (28%) do not know about the IT GRC program in their organizations. These participants were not included in the analysis of the results.

Among organizations that implemented the IT GRC program, 71.9% of the participants stated that their organizations have an integrated IT GRC program that covers all of governance, risk management, and compliance processes. 77.4% of the participants also reported that the person in charge of IT GRC in the organization is from the IT department (i.e., Chief Information Officer (CIO) or equivalent or another IT director). Table III shows IT GRC status in Saudi healthcare organizations.

C. Reliability Testing

Cronbach's coefficient alpha test was used to measure internal consistency between items of the contexts. All of the values of Cronbach Alpha were above 0.8, which is considered above the accepted threshold [28].

TABLE II. THE DEMOGRAPHIC CHARACTERISTICS OF THE PARTICIPANT'S CHARACTERISTICS

Characteristic	Total Respondents	
	Frequency	Percentage
Gender		
Female	56	47.5
Male	62	52.5
Total	118	100
Organization location		
Central Province	78	66.1
Western Province	30	25.4
Eastern Province	6	5.1
Southern Province	2	1.7
Northern Province	2	1.7
Total	118	100
Participants' experience in years in healthcare sector		
Less than one year.	26	22.0
More than 1 year and less than 5 years.	13	11.0
More than 5 years and less than 10 years	19	16.1
More than 10 years and less than 15 years.	25	21.2
15 years or more.	35	29.7
Total	118	100
Participants' experience in years in IT GRC-related activities		
Less than one year.	65	55.1
More than 1 year and less than 5 years.	28	23.8
More than 5 years and less than 10 years	17	14.4
More than 10 years and less than 15 years.	7	5.9
15 years or more.	1	0.8
Total	118	100

Overall findings

This research aims to study the factors affecting the implementation of IT GRC initiatives in healthcare organizations in Saudi Arabia. The authors have addressed this by analyzing the data collected from the survey. Among the six contexts, the most important one is Legal (mean 3.85), then Processes (mean 3.82), followed by Strategy (mean 3.78), Technology (mean 3.74), Business (mean 3.62), and finally People (mean 3.61). Table IV shows the overall results of the analysis.

D. Comparison between Different Groups

Based on the results of the questionnaire, Saudi healthcare organizations divide into three categories. The first category is the organizations that have already utilized IT GRC solutions. The second category is the organizations that are planning to implement IT GRC programs. The third category is the organizations that do not implement the IT GRC program. The researchers conducted an Analysis of Variance (ANOVA) test to examine the means among the three groups. Table V lists the

means and standard deviations for all groups and contexts. It also presents the p-value for the ANOVA test for all contexts among the three groups. The results indicated significant differences among all six contexts between all categories. These findings show that organizations with IT GRC programs always have high mean values for all contexts. Additionally, organizations that do not implement IT GRC programs always have low mean values for all contexts. The ranking of the mean values shows differences in the importance of contexts for each category. For organizations with an IT GRC program, the processes context ranked first (mean = 4.10), and the people context was ranked last (mean= 3.89). For organizations planning to implement IT GRC programs, the legal context was ranked first (mean = 3.78), and the business context was ranked last (mean= 3.33). For organizations that do not implement IT GRC programs, the legal context was ranked first (mean = 2.77), and the process context was ranked last (mean= 3.22).

TABLE III. IT GRC STATUS IN SAUDI HEALTHCARE ORGANIZATIONS

Question	Total Respondents	
	Frequency	Percentage
Having an IT GRC program in place?		
The organization has IT GRC program.	57	48.3
The organization is planning to have IT GRC program in the foreseeable future.	19	16.1
The organization does not have any IT GRC program.	9	7.6
I do not know.	33	28.0
Total	118	100
Type of IT GRC program		
The organization has an integrated IT GRC program that covers all of governance, risk management, and compliance processes.	41	71.9
The organization has an IT GRC program that covers only two aspects of GRC processes (i.e., governance and risk management, governance and compliance, or risk management and compliance).	11	19.3
The organization has an IT GRC program that covers only one aspect of GRC processes (i.e., governance, risk management, or compliance).	5	8.8
Total	57	100
Roles leading IT GRC program in the organization?		
Chief Information Officer (CIO) or equivalent	32	56.4
Deputy CIO or equivalent	3	5.2
Chief Cybersecurity officer (CCO) or equivalent	9	15.8
Another IT director/manager	9	15.8
Internal audit officer.	2	3.4
Other	2	3.4
Total	57	100

TABLE IV. OVERALL ANALYSIS

Context	Rank	Mean	SD	Number of items	Cronbach Alpha
Legal	1	3.85	1.06	4	0.955
Process	2	3.82	1.08	4	0.950
Strategy	3	3.78	1.12	4	0.948
Technology	4	3.74	0.972	5	0.904
Business	5	3.62	1.21	4	0.956
People	6	3.61	1.19	2	0.893

VI. DISCUSSION

The present research tries to explain the factors affecting the implementation of IT GRC programs in Saudi healthcare organizations. Legal context, which refers to the ability of healthcare organizations to meet all legal and regulatory requirements, was ranked as the most important context. This result was expected since following legal and regulatory requirements is usually compulsory. However, organizations that do not implement IT GRC programs face difficulties following or complying with these requirements since they have low mean values (2.72). This finding aligns with other studies that indicate compliance with the laws and regulations as one of the main objectives for GRC implementation [24].

The process dimension that controls various GRC domains ranked as the second most important dimension. Our study shows that organizations with well-structured processes have already implemented IT GRC since this dimension was ranked first for such organizations. This dimension ranked as the last dimension for organizations without an IT GRC implementation. Thus, they have difficulties controlling IT risk management and IT compliance management. Another study supports this finding since it showed the process level of the integration between IT GRC domains was low at healthcare organizations compared with other industries [23].

Synergy and alignment between IT GRC activities, IT strategic objectives, and business strategy are among the most crucial conditions. For organizations that plan to implement IT GRC solutions, this context ranked as the second most important. This ranking shows their willingness to offer better alignment integration between the GRC processes, IT, and business strategies. Weak strategic alignment negatively impacts IT GRC integration efforts across the healthcare organizations as mentioned in [1].

Technology context ranked as the 4th most important context. One possible reason is the use of technology as a tool to support IT GRC implementation rather than the main focus for healthcare organizations [23]. Another possible reason is the technical complexity of implementing such solutions in healthcare organizations [24].

Business context refers to the financial issues regarding the implementation of IT GRC activities in the organization. This context did not rank high in our study. This result was a surprise since financial factors are among the most important factors in many studies related to the implementation of IT in healthcare organizations [29]. A possible explanation is continuous pressure on healthcare organizations to decrease their expenditures [30].

TABLE V. CONTEXTS ACROSS DIFFERENT GROUPS

Context	Having an IT GRC program		Planning to have an IT GRC program		Do not implement an IT GRC program		P value
	Mean	SD	Mean	SD	Mean	SD	
Legal	4.05	1.009	3.78	0.942	2.72	0.947	.004*
Process	4.10	0.939	3.74	0.963	2.22	0.785	<.001*
Strategy	3.96	1.049	3.75	1.118	2.67	1.053	.011*
Technology	3.99	0.965	3.53	0.706	2.64	0.615	<.001*
Business	3.91	1.106	3.33	1.294	2.39	0.719	<.001*
People	3.89	1.005	3.39	1.329	2.28	1.034	.001*

* Statistically Significant

People context includes human involvement in IT GRC activities in healthcare organizations. This context ranked with low mean values. The reason could be the lack of adequate staff devoted to IT GRC activities [1]. Another explanation is the unclear responsibilities for the IT GRC team [31]. Our finding also indicated a lack of expertise in the IT GRC domain in healthcare since only 21.1% of the participants have more than five years of experience in IT GRC related activities. This result shows the need for specialist training in the IT GRC domain in Saudi healthcare organizations.

VII. CONCLUSION

Healthcare organizations face many challenges to improve their operations to respond to regulatory, strategic, and other requirements. Many organizations have implemented the Governance, Risk, and Compliance (GRC) model to help manage and comply with internal and external legal aspects. IT GRC is a subdomain of GRC that focuses on IT operations in organizations. This study analyzed the factors affecting the implementation of IT GRC in healthcare organizations. It developed the research framework that comprises strategy, processes, technology people, and legal and business requirements.

The results indicated significant differences among all six contexts between various categories of healthcare organizations. The output of our research provides a strategic roadmap for healthcare organizations that are willing to implement IT GRC activities. Additionally, Saudi healthcare organizations need to pay special attention to the role of people in IT GRC activities since this context ranked among the lowest.

The main contribution of this study is to develop a strategic framework for IT GRC in healthcare organizations. The developed framework can help healthcare organizations in improving their IT services and align them with healthcare services. Another implication of this study is the need to align technology with other aspects such as legal requirements for IT GRC holistic strategic framework.

A possible limitation of the research is the number of participants, which can be considered low. However, the low number could be because our survey requires participants to have some knowledge of IT GRC to complete the questionnaire, and these people are usually limited in

healthcare organizations. Another limitation is that the geographical context of the study is limited to Saudi Arabia.

Future work of our study includes the development of tools that integrate all the contexts to support healthcare organizations for better utilization of IT GRC concepts.

REFERENCES

- [1] M. Krey, "Significance and current status of integrated IT GRC in health care: An explorative study in swiss hospitals," in Proceedings of the Annual Hawaii International Conference on System Sciences, 2015, vol. 2015-March, pp. 3002–3012. doi: 10.1109/HICSS.2015.363.
- [2] N. Racz, E. Weippl, and A. Seufert, "A Frame of Reference for Research of Integrated Governance, Risk and Compliance (GRC)," Dec. 2010, vol. 6109, pp. 106–117. doi: 10.1007/978-3-642-13241-4_11.
- [3] G. E. Handoko BL, Riantono IE, "Importance and Benefit of Application of Governance Risk and Compliance Principle," SRP, vol. 11, no. 9, pp. 510–513, 2020.
- [4] F. Lezzar, D. Benmerzoug, K. Ilham, and A. Osmani, "A GRC-Centric Approach for Enhancing Management Process of IoT-Based Health Institution," in Lecture Notes in Computer Science, 2017, pp. 207–221. doi: 10.1007/978-3-319-67807-8_16.
- [5] R. R. Moeller, "Importance of Governance, Risk, and Compliance Principles," in COSO Enterprise Risk Management, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011, pp. 21–29. doi: 10.1002/9781118269145.ch2.
- [6] N. Racz and A. Seufert, "A.: A frame of reference for research of integrated governance, risk & compliance (GRC)," pp. 106–117, 2014.
- [7] T. Greenhalgh, "Aligning Governance, Risk, and Compliance," Ahima, 2020. <https://journal.ahima.org/aligning-governance-risk-and-compliance/> (accessed Nov. 20, 2021).
- [8] C. A. Hardy and J. Leonard, "Governance, risk and compliance (GRC): Conceptual muddle and technological tangle," 2011.
- [9] M. Nicho, S. Khan, and M. S. M. K. Rahman, "Managing Information Security Risk Using Integrated Governance Risk and Compliance," in 2017 International Conference on Computer and Applications (ICCA), Sep. 2017, pp. 56–66. doi: 10.1109/COMAPP.2017.8079741.
- [10] T. Haugh, "Harmonizing Governance, Risk Management, and Compliance through the Paradigm of Behavioral Ethics Risk," Corporate Law: Corporate Governance Law eJournal, 2019.
- [11] A. Appari and M. Johnson, "Information Security and Privacy in Healthcare: Current State of Research1," International Journal of Internet and Enterprise Management, vol. 6, pp. 279–314, 2010, doi: 10.1504/IJEM.2010.035624.
- [12] A. C. Cagliano, S. Grimaldi, and C. Rafele, "A systemic methodology for risk management in healthcare sector," Safety Science - SAF SCI, vol. 49, pp. 695–708, 2011, doi: 10.1016/j.ssci.2011.01.006.
- [13] M. Krey, "Information technology governance, risk and compliance in health care - A management approach," Proceedings - 3rd International Conference on Developments in eSystems Engineering, DeSE 2010, pp. 7–11, 2010, doi: 10.1109/DeSE.2010.8.
- [14] H. Abdullah, "Analyzing the technological challenges of Governance, Risk and Compliance (GRC)," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), pp. 274–282, 2019.
- [15] A. M. H. Kuo, "Opportunities and challenges of cloud computing to improve health care services," Journal of Medical Internet Research, vol. 13, no. 3, 2011, doi: 10.2196/jmir.1867.
- [16] M. S. Ph. D. Murray Peter J. and R. J. Ward, "Promoting Enterprise Risk Management (ERM) and Governance, Risk and Compliance (GRC) for Managing Cybersecurity Risks." 2018.
- [17] R. Batenburg, M. Neppelenbroek, and A. Shahim, "A maturity model for governance, risk management and compliance in hospitals," Journal of Hospital Administration, vol. 3, Feb. 2014, doi: 10.5430/jha.v3n4p43.
- [18] M. Krey, "Significance and Current Status of Integrated IT GRC in Health Care: An Explorative Study in Swiss Hospitals," 2015 48th Hawaii International Conference on System Sciences, pp. 3002–3012, 2015.

- [19] C. L. Wilkin and R. H. Chenhall, "Information Technology Governance: Reflections on the Past and Future Directions," *Journal of Information Systems*, vol. 34, no. 2, pp. 257–292, Oct. 2019, doi: 10.2308/isy-52632.
- [20] Greenhalgh, "Aligning Governance, Risk, and Compliance," *JOURNAL of AHIMA*, 2020.
- [21] M. Nicho, S. Khan, and M. S. M. K. Rahman, "Managing Information Security Risk Using Integrated Governance Risk and Compliance," 2017 International Conference on Computer and Applications, ICCA 2017, pp. 56–66, 2017, doi: 10.1109/COMAPP.2017.8079741.
- [22] P. Vicente and M. M. da Silva, "A Business Viewpoint for Integrated IT Governance, Risk and Compliance," in 2011 IEEE World Congress on Services, 2011, pp. 422–428. doi: 10.1109/SERVICES.2011.62.
- [23] N. Racz, E. Weippl, and R. Bonazzi, "IT Governance, Risk & Compliance (GRC) Status Quo and Integration: An Explorative Industry Case Study," 2011, pp. 429–436. doi: 10.1109/SERVICES.2011.78.
- [24] A. Papazafeiropoulou and K. Spanaki, "Understanding governance, risk and compliance information systems (GRC IS): The experts view," *Information Systems Frontiers*, vol. 18, no. 6, pp. 1251–1263, 2016, doi: 10.1007/s10796-015-9572-3.
- [25] N. Racz, E. Weippl, and A. Seufert, "A Frame of Reference for Research of Integrated Governance, Risk and Compliance (GRC)," in *Lecture Notes in Computer Science*, B. de Decker and I. Schaumüller-Bichl, Eds. Springer, 2010, pp. 106–117. doi: 10.1007/978-3-642-13241-4_11.
- [26] Jamovi, "The jamovi project." Jamovi, 2021. [Online]. Available: <https://www.jamovi.org>.
- [27] MoH, "Statistical Yearbook 2020," 2021. <https://www.moh.gov.sa/en/Ministry/Statistics/book/Pages/default.aspx> (accessed Nov. 15, 2021).
- [28] M. Krey, T. Keller, B. Harriehausen, and M. Knoll, "Towards a classification of Information Technology governance frameworks for the development of a IT GRC healthcare framework," 2011 IEEE Consumer Communications and Networking Conference, CCNC'2011, pp. 34–38, 2011, doi: 10.1109/CCNC.2011.5766488.
- [29] F. Alharbi, A. Atkins, and C. Stanier, "Understanding the determinants of Cloud Computing adoption in Saudi healthcare organisations," *Complex & Intelligent Systems*, vol. 2, no. 3, pp. 155–171, Oct. 2016, doi: 10.1007/s40747-016-0021-9.
- [30] A. O. Babatunde, A. J. Taiwo, and E. G. Dada, "Information Security in Health Care Centre Using Cryptography and Steganography," *arXiv*, vol. 14, no. 2, pp. 172–182, 2018.
- [31] A. Papazafeiropoulou and K. Spanaki, "Understanding governance, risk and compliance information systems (GRC IS): The experts view," *Information Systems Frontiers*, vol. 18, no. 6, pp. 1251–1263, Dec. 2016, doi: 10.1007/s10796-015-9572-3.

Ambulatory Monitoring of Maternal and Fetal using Deep Convolution Generative Adversarial Network for Smart Health Care IoT System

S. Venkatasubramanian

Department of Computer Science
Saranathan College of Engineering, Trichy-620012, India

Abstract—With the increase in the number of high-risk pregnancies, it is important to monitor the health of the fetus during pregnancy. Major advances in the field of study have led to the development of intelligent automation systems that enable clinicians to predict and determine the monitoring of Maternal and Fetal Health (MFH) with the aid of the Internet of Things (IoT). This paper provides a solution for monitoring high-risk MFH based on IoT sensors, data analysis-based feature extraction, and an intelligent system based on the Deep Convolutional Generative Adversarial Network (DCGAN) classifier. Various clinical indicators such as heart rate of MF, oxygen saturation, blood pressure, and uterine tonus of maternal are monitored continuously. Many data sources produce large amounts of data in different formats and ratios. The smart health analytics system proposes to extract several features and measure linear and non-linear dimensions. Finally, a DCGAN has been proposed as a predictive mechanism for the simultaneous classification of MFH status by considering more than four possible outcomes. The results showed that the proposed system for mobile monitoring between MFH is a practical solution based on the IoT.

Keywords—Deep convolutional generative adversarial network; fetal health monitoring; high-risk pregnancies; internet of things; smart healthcare system

I. INTRODUCTION

Remote monitoring systems in the healthcare domain are increasing the daily reach of health for at-risk populations, especially pregnant women [1] and the elderly [2]. To identify the early disease symptoms and provide care, the patients are monitored every second by using these promising techniques in healthcare. The major functionality of the system is to diagnose and predict the health conditions of the user and provide warnings and training for the same. Nowadays, recent advances in the technologies of IoT have presented a way to enable such monitoring services 24/7. The IoT is a growing network of interconnected objects that include shared knowledge about decision-making and efficient and autonomous operation [3-6]. Various sources such as computer knowledge, communication link, and sensitivity are used by IoT in healthcare. As the content of public health, MFH is highly regarded by governments. However, medical care services, especially obstetric care, are limited, which decreases the efficiency of the medical staff and increased the pressure of high-quality service.

Although the fetus is located inside the body of a pregnant woman, the most devoted protector of the fetus is only the pregnant woman. Without the help of external technology, a mother will not be able to know the FH even when it is life-threatening. Therefore, to study the MH, monitoring the fetal is an important tool [7]. Fetal monitoring is reliable, safe, and easy to operate, which is widely practiced by MCH organizations across India [8-10]. However, the fetal monitoring used by most hospitals still traces deficiencies. First, the information cannot be shared. Monitor results must be printed in most hospitals, making them unsuitable for storage and easy to lose, and sharing tracking information and advice from multiple individuals may not feel timely.

These factors make it difficult for doctors to perform medical tasks or delay illness, which can lead to medical accidents. When pregnant women are in the obstetrics department, early warning function and intelligent real-time monitoring system are unable and they face other issues such as high risk, high emergencies, and massive flow of other pregnant women. In addition, serious consequences and life-threatening risks have occurred, when pregnant women are unable to notify doctors in time. Finally, monitoring of human health at home fails to recognize: at present, traditional health systems cannot meet the needs of continuous monitoring of pregnant women. Pregnant women, as expectant mothers, play an indispensable role in the upbringing of their children. Therefore, self-monitoring at home is an important content during the period of perinatal health care.

In recent years, communities and families are greatly entering into the comprehensive application of information platforms such as cloud computing technology, IoT healthcare, big data, communication technology, etc. It provides individuals with purposeful and personalized services, enabling them to advance from telemedicine to new ideas and ways of preventing disease and addressing major public health issues in the lives of women and children [11-13]. The IoT [14] means things are connected via the Internet that, understands the communication between objects using modern information technologies such as smart sensing, identification technology, and wireless communication. IoT-based M-Health uses wearable clinical sensors [15-17] to provide patients with real-time feedback on key symptoms and medical information and provide them with "anytime, anywhere" healthcare.

Healthcare as a Service platform architecture for cloud-based medical decision support services is developed in [18]. In [19], a coherent framework for M-Health monitoring and IoT-based remote monitoring is proposed. In [20], the IoT-based cloud service is optimized for the next generation of smart environments. In wireless heterogeneous networks, architecture of hierarchical sensor-based healthcare is developed by the work [21]. Other work [22] focused on a variety of wireless access networks and provided a framework for remote patient monitoring services. In some other cases, the software is developed by the IoT infrastructure for safe and smart healthcare [23].

An example of GAN model architecture is shown as below in Fig. 1.

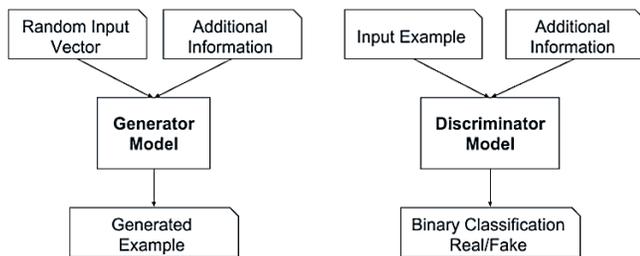


Fig. 1. Example of Generative Adversarial Network Model Architecture.

An integrated solution is proposed in this work for collecting different data from devices and sensors of IoT, linear and non-linear features are extracted, emergency alarms are used to analyze the data and finally, the MFH is automatically predicted by the DCGAN model. The most significant contributions are presented here as:

- 1) A continuous monitoring of MFH is carried out by integrating the medical devices and sensors of IoT.
- 2) To improve the diagnostic accuracy, fog computing architecture is used to develop an automated emergency subsystem with inference rules set.
- 3) The extraction of linear and non-linear features are done by a smart health analytics system.
- 4) The status of MFH is classified by analyzing the multiple metrics with the help of the DCGAN model.

This research paper is structured as follows. The related works for monitoring the fetal and maternal status are given in Section 2. A brief explanation of the system model with a predictive model is presented in Section 3. The validation of existing techniques with the proposed algorithm is described in Section 4. Finally, the scientific contribution of the research work with its future development is depicted in Section 5.

II. RELATED WORK

A safe and reliable monitoring framework at less cost-effective is developed by Allah et al., [24] to make the home more comfortable and mitigate the effects of preterm labor in pregnant women. Non-invasive technology is used by the system to monitor EHG defects using wireless body sensors and smartphones. The smartphone will check for contractions and alert you in case of early delivery. Using the cervical

contractility database, this smartphone app has been developed and tested for verifying reliability and performance in terms of power consumption. The analysis showed that this application serves the purposes of the framework in defining the work system.

The significance of variants with similar features between the EPL models and the current pregnancy models is proposed by Liu, et al., [25]. After the collection of embryo samples, the correlation between the heart rate of the fetus is identified, where a regression model is used to achieve the normal development of FHR. The remaining analysis reveals the importance of the FHR in determining pregnancy outcomes. To develop the computational models, this paper developed six different machine learning (ML) techniques are developed. Sensitivity is used to compare the accuracy of both the presence of FHR and conditions of absence for predicting the performance. FHR is closely related to ETD for normal development and attained a high-performance value. When compared with all techniques, the random forest has 97% of recall, F1-score, and accuracy with 0.97 of AUC; however, deep learning techniques are required for effective prediction.

During pregnancy, the exposure of incarceration and food insecurity is identified by Testa et al., [26]. From 2004 to 2015, the LR was used to determine the relationship between father and mother of a fetus' food insecurity using Pregnancy Risk Assessment System (PRAMS) results. Withdrawal of controls is associated, directly or indirectly, with a 165% increase in the risk of food instability. Attendance analysis indicates that this relationship is driven by various factors such as receiving WIC benefits, unemployment of maternal and financial hardship.

Azimi et al., [27], established a flexible decision-making mechanism to provide 24/7 health outcomes in the absence of data, despite missing policies. Various data resources are leveraged in IoT systems for providing the results and imputing the missing values. This approach was validated in the Human Maternal Health Cycle Research, where 20 pregnancies were observed for 7 months. The health status of the maternal is measured by using her heart rate in real-time applications. While comparing with existing related works, the accuracy is highly achieved by this developed method. But, the accuracy results are low, when there are large missing window size data presents.

The website is developed by Mourad M et al., [28], and the signalling site is used to study uterine contractions, their characteristics, and labour study. In this investigation, different nonlinear techniques are used to determine pregnancy and labor signals and to examine pregnancy signals before labor with Hogworth frontiers, the multifaceted nature of Lempel-Ziv, and the measurement of a fractal. In Lebanon and France, all data from 12 WBL to 1 WBL are recorded and played back using a terminal network of size 4×4 . The engineered signals are used by these techniques for testing the sensitivity of nonlinearity modifications, which are then applied to real signals. These results show that nonlinear techniques are commonly used to detect pregnancy fluctuations and to reduce symptoms.

III. PROPOSED SYSTEM

Fig. 3 shows an overview of the proposed structure. The flow of the methodology is depicted as follows: initially, the IoT devices are used to generate the data and then transmitted to the emergency subsystem for identifying any distress of FHS. The medical staff gets informed if any emergency is obtained. After all this, data is transferred to the cloud, where the calculation of features is processed and a prediction system based on one-dimensional DCGAN is used. At last, medical diagnoses are supported by this classification model and the results are provided to the clinical staff members for further analysis.

The medical staff only has access to patients' details and their diagnostic outcomes, where authentication, access control layer, and data encryption levels are used to provide communication protection for data privacy and confidentiality.

There are four parts in the developed solution, which are presented in Fig. 2. The integration of medical devices and sensors for the collection of data is presented in the first layer called the IoT layer, where the emergency subsystem is presented in the second layer (i.e. Fog computing layer) that is according to the inference rules and fixed clinical thresholds. The third and final part is considered as the Cloud computing layer, where signal processing techniques are used to extract the features in the third part, and classification of MFH is carried out by one-dimensional DCGAN in the fourth layer.

A. IoT Layer

Sensor and Equipment Devices which are used to collect medial parameters from mother and fetal.

The components that are presented in this layer are described as: Four vital signs of maternal such as blood

pressure, temperature, heart rate of both Mother and fetal and oxygen saturation. Fetal and maternal monitoring are developed by using IoT modules, where diagnostic support is also provided by non-structured data.

To continuously monitor the fetus, the heart rate of his/her is only considered, which is obtained by the sensor of Doppler with a 4 Hz sampling rate. Then, a toco-dynamometer sensor is used to monitor the uterine tonus activity of maternal with the same sampling rate. Fig. 4 shows the modelling of the IoT interface layer.

At a 1 Hz sampling rate, important signals of maternal are obtained, where a photoplethysmography sensor is used to derive the oximetry and heart rate of maternal; then digital sensor integrated with medical devices is used to acquire the blood pressure for systolic and diastolic with temperature. Then, relevant events, audio, and photo notes of maternal that are acquired by medical staff are shared with the specialists, who are added to this system.

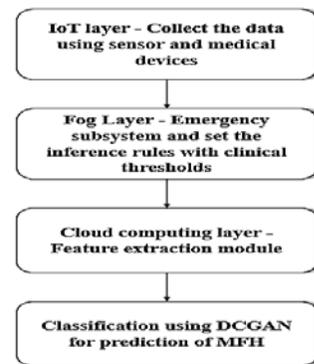


Fig. 2. The Flow of the Proposed Structure.

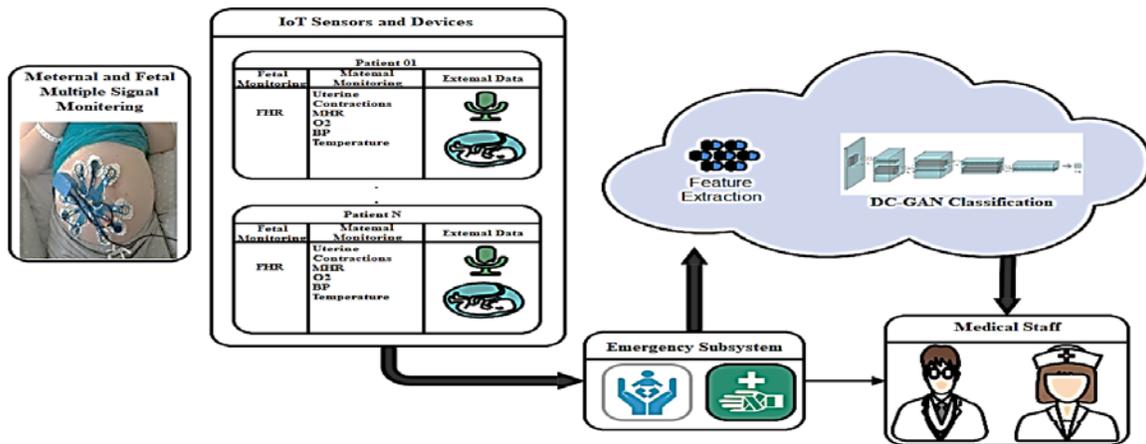


Fig. 3. Proposed Healthcare-IoT for Simultaneous Monitoring of a Mother and Fetus.

Patient 01		
Fetal Monitoring	Maternal Monitoring	External Data
FHR	Uterine Contractions MHR O2 BP Temperature	 

Fig. 4. Representation of Patient's Data with the Generation of non-external Data.

B. Fog Computing Layer: Subsystem for Emergency

With limited resources of technology in remote clinical facilities and developing countries for providing ambulatory support, management and adoption of a cloud computing system for monitoring the real-time data is a challenging task. Therefore, architecture on fog computing is developed for supporting the requirements by creating a layer to collect the data on clinical premises.

For the signals of UC and heart rate of fetal, this type of external sensor is ideal for sound generation depending on the situation or the mother's movements [29]. Essentially, a pre-processor is included in the model before considering the specific dimensions of noise filtering and zero detection. The selected layer develops an automated analytical system for clinically important signals, according to the set of defined limits. The system is designed using three types of diagnostics (i.e. Emergency Class as EC) given in Table I.

According to the International Guidelines for Obstetricians and Gynaecologists (IGOG) [30], detailed descriptions of each standard portal and related classifications are provided in Table II for guidance on clinical group definitions and parental criteria.

C. Cloud Computing Layer: Subsystem for Extraction of Automatic Features

A data analytical module is created in this part for the extraction of maternal and fetus' features. To determine the parameters of UC and heart rate of fetus, signal processing techniques [31] are used and then non-linear and statistical metrics are calculated. The changes include long decelerations and baseline adjustments that need time to confirm and long-term analysis are required by interpretation of fetal's heart rate. Therefore, once the signal acquisition is done, after 10 minutes only the measures are calculated at the initial step and updated every 5 minutes for the entire examination.

There are two steps involved in this module, the initial importance is given for monitoring the fetal and finally, the vital signs of the mother has considered as second steps. For continuous monitoring, maximum information is achieved by analyzing the uterine contractions and heart rate of fetal by using signal processing techniques. In this sub-system, a total of 15 features are extracted for the analysis of the heart rate of the fetus that is established by medical experts, where the features include baseline value, baseline changes for instant and numbers, minima number, FHR minima instants, peaks for number and instants, decelerations of DIP-I and DIP-II, UC occurrences for number and instants, sample entropy for 5 and 20 minutes window, long decelerations and variable decelerations.

D. Predictive Subsystem using DC-GAN Model

GANs (Generative Adversarial Networks) are made up of two models that are trained concurrently using an adversarial approach. As shown in Fig. 5(a) and Fig. 5(b), a generator learns to make realistic images, whereas a discriminator learns to distinguish between actual and fraudulent images.

It should be noted that a total of 6 classes are considered and included in Table III for maternal and fetal diagnostic

diagnoses based on medical group classification. In this work, a DC-GAN is used for the prediction process.

In this paper, the conditional DC-GAN is used to generate conditional deep changes according to DC-GAN [32] and cGAN [33]. In the original DC-GAN architecture, the bias was made up of step confusion layers, modular correction layers, and LeakyReLU activations.

The generator contains spasms, translucency, patches, and ReLU activations. From part 3 of the cloud computing layer, the input of the selected classifier consists of 15 features of the system to predict the six possible outputs, which are described in Table III. Fig. 5 shows the configuration of the DC-GAN state.

Instead of the original layer of convolutional-transpose, an up-sample and convolutional layer are used in a generator for avoiding the artifacts of checkerboard [34]. If standard deconvolution is used to scale up from low to high precision, it uses all of the small feature points to 'plot' a larger square. This is where the "unequal overlap" problem comes in, i.e. when these "squares" converge on the larger side. In particular, D convolutions are random nests when the size of the kernel is not divisible by a step. Overlapping pixels create an unnecessary chessboard shape on the generated images. So, to get the "de-Convolution" function, we used the sample layer above instead. We scaled the feature using the top sample layer using bilinear (or adjacent interpolation) and then the convolutional layer. To achieve better results, a RelectionPad2d layer with size 1 is added before the convolutional layer for avoiding the boundary artifacts.

TABLE I. EMERGENCY SUBSYSTEM OUTPUT'S CLASSES

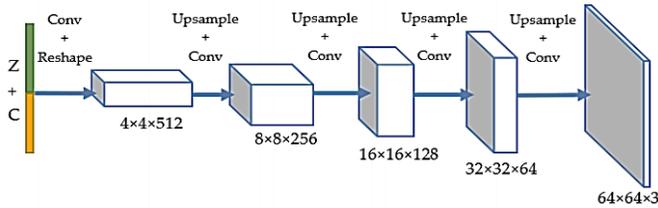
Layer's Outcome	Explanation
EC1 as 1	Fetal Emergency (FE)
EC2 as 2	Maternal Emergency (ME)
EC3 as 3	Maternal and Fetal Emergency (FME)

TABLE II. LIST OF EMERGENCY LIMITS/THRESHOLDS BASED ON THE ABOVE GUIDELINES

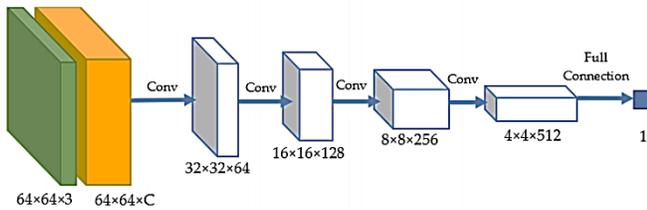
Clinical Parameter	Threshold	Interpretation	Output Class
Heart Rate of Fetus	> 160	Normal tachycardia	1
	> 180	Severe tachycardia	1
	< 100	Normal bradycardia	1
	< 80	Severe bradycardia	1
Heart Rate of Mother	> 100	Tachycardia	2
	< 60	Bradycardia	2
Oximeter	< 90	Mother's hypoxemia	2
Temperature	> 37.5	Fever for mother	2
Systolic blood pressure	> 140	Mother blood pressure	2
Diastolic blood pressure	> 90	Blood pressure for mother	2

TABLE III. PREDICTED CLASSES FOR THE PROPOSED SUBSYSTEM

Output	Description
PC1	Normal for both Maternal and Fetal
PC2	Suspicious only on fetal
PC3	Distress on fetal
PC4	Suspicious only on maternal
PC5	Harmful for Maternal
PC6	Harmful for both



(a).Generator.



(b).Discriminator.

Fig. 5. Structure of Conditional DC-GAN.

To control the output function, the input model gets a class designation for the generator and the amplifier. The generator receives a random beep ($1 \times 1 \times Z$) and a heat conditional sign ($1 \times 1 \times C$), where the dimensions Z and C represent the number of categories. By default, we chose $Z = 100$ and $C = 60$. They were combined and sent to the curved layer using a 1×1 core. Output channel 8192 ($= 4 \times 4 \times 512$). Later, the output data was converted to $4 \times 4 \times 512$ format. Finally, we sent the data to several modules (except for Up-sample + Conv2D + Batch-Norm, the output layer) as high as the original DC-GAN. An equal-valued data is up-sampled, only when the data size is 1×1 , and therefore, input data i.e. concatenation of one-hot label and random noise is not resized directly. Hence, the input noise size is kept at $1 \times 1 \times Z$. To make the resizing step work, an extra convolutional layer is added in this model, where the C channel is encoded with a label in the discriminator. More precisely, the function label n is the data of the $H \times W \times C$ scale, each 1s and 0.s in the n channel. In the input layer, the feature is associated with the label: each feature is associated with the label. Channel direction label. Thus, the input channel size was $3 + C$. Similar to the original DC-GAN configuration, the Discriminator configuration has multiple configuration layers and a default volume layer (except for the output layer).

E. Training Process and Loss Function

We used the standard loss function for GANs, which is defined in [35] as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

There are two losses such as $Loss_D$ and $Loss_G$ used since a generator and a discriminator are presented in the GAN model. Here, the loss function is used as Binary Cross-Entropy (BCE).

A sum of $errD_{real} = -\log(D(x))$ and $errD_{fake} = -\log(1 - D(G(z)))$ is used to calculate the first loss, where the BCE is represented to the reality only when the fake inputs and real inputs are received by D . The prediction of false positive is carried out by calculating the BCE when the fake inputs are received by D , i.e. $Loss_G = -\log(D(G(z)))$.

Set the default value for Adam optimizer as learning rate = 0.0001, when the training and updating of the generator and discriminator are carried out one by one. Finally, six possible outcomes are derived and the next section will show the validation of the proposed model with existing systems.

IV. VALIDATION AND RESULTS

In this section, details of the datasets used in testing evaluation for a particular system are provided. Presenting the results obtained with the contingency subsystem, the following section validates the predictive subsystem feature of DCGAN with or without extrusion, using existing deep learning techniques. The performance measures used in this study are described in this section.

A. Dataset Description

Simulation experiments were performed on a standard antenatal CTG database from the UCI ML Repository (University of California, Irvine) [36]. CTG data were collected from 29 to 42 weeks gestation at SisPorto 2.0 in Portugal [37]. The database contains 2126 cases with 21 traits and cases classified according to FIGO guidelines by the consensus of three experts: 1655 general cases, 295 suspicious, and 176 pathological data.

B. Performances Metrics

In this section, major parameter metrics are discussed to validate the system model's performance. The correct identification of data samples over a total number of instances is calculated by using accuracy. An identification of MFS is often accurate by using the parameter called precision or positive predictive value. The possible outcomes are correctly classified by the proposed model is known as recall/sensitivity. The proportion of actual negatives is calculated by using the true negative rate/specificity. The harmonic average of sensitivity and precision is provided by F1-score or F-measure.

Accuracy is calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision is working out as following

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Sensitivity is deliberate as follows.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

Specificity and F-measure are defined as follows:

$$Specificity = TN/(FP + TN) \text{ or } (1 - FPR) \quad (5)$$

$$F1 - score = 2TP/(2TN + FP + FN) \quad (6)$$

C. Evaluation of Proposed Emergency Subsystem

All emergencies have been classified by the organization. The results in Table IV show that the sensitivities (Caesarean deliveries as CI) for all classifiers are very low and the associated features are high. This is expected because the database is skewed in favor of genital records. Sensitivity and 95% adjusted specificity for AUC and MSE are determined using the recommended CI model.

Disagreements during the modelling phase may result in poor performance. The integrating parameters from Table II are used to add the inference rule set to the system for improving the accuracy levels. For instance, the emergency of MFH is highly calculated by diagnosing the elevated blood pressure for systolic and diastolic, the temperature of maternal, etc. After discussing with the medical team, a total of twenty-six hypothetical rules are generated and here, five samples are provided in Table V.

The performance of the emergency subsystem based on the set of standard limits is shown in the Table VI.

It is important to note that performance is unsatisfactory with FP and FN rates high and sensitivities below 80%, which is unacceptable given the critical interpretation of the contingency. The new set of proposed default rules has slightly improved the performance of the contingency.

TABLE IV. PERFORMANCE OF THE EMERGENCY SUBSYSTEM

Output	Performance of Model before setting the static threshold			
	Sensitivity	Specificity	AUC	MSE
ME	0.00(0.00,0.00)	0.99(0.99,0.99)	0.60(0.58,0.61)	0.08(0.07,0.08)
FE	0.02(0.01,0.03)	0.99(0.99,0.99)	0.68(0.65,0.69)	0.08(0.07,0.08)
FME	0.02(0.00,0.04)	0.99(0.99,0.99)	0.71(0.68,0.73)	0.08(0.07,0.08)

TABLE V. THE SUBCOMMITTEE ON CONCEPTUAL LAWS CONSIDERED TO DESCRIBE THE EMERGENCY FOR MOTHER OR FETUS

Rule	Output
[MHR > 90] and [blood pressure for systolic > 120]	ME
[Temperature < 36] and [blood pressure for systolic < 100] and [blood pressure for diastolic < 60]	ME
[MHR > 80] and [FHR > 140]	FE
[FHR > 150] and [MT > 37]	FE
[blood pressure for diastolic > 13] and [blood pressure for systolic > 13] and [MHR > 90] and [FHR > 120]	MFE

TABLE VI. EMERGENCY SUBSYSTEM PERFORMANCE UNDER EXPERT SUPERVISION AND DEFAULT RULES SET WITH FIXED LIMITS

Output	Considering the set of static thresholding			
	Sensitivity	Specificity	AUC	MSE
ME	0.76(0.70,0.81)	0.56(0.54,0.58)	0.70(0.68,0.73)	0.08(0.07,0.08)
FE	0.71(0.68,0.73)	0.82(0.80,0.85)	0.87(0.86,0.88)	0.18(0.17,0.19)
FME	0.87(0.85,0.88)	0.91(0.89,0.92)	0.96(0.96,0.97)	0.08(0.07,0.09)

The new set of proposed default rules has slightly improved the performance of the contingency subsystem. By considering only the maternal prognosis, the AUC (ME) is 70%. For fetal emergency diagnosis, the system performance is AUC (FE) 87%. This can happen because there is only one parameter that is directly related to the condition of the fetus, which is the FHR. Finally, in the worst-case scenario, when the mother and fetus have an emergency, the specific regimen performs best, with an AUC (FME) of 96%.

D. Performance Analysis of DC-GAN Prediction Subsystem

In this section, five parameters are used to validate the proposed DC-GAN with existing techniques by considering with and without feature extraction techniques. and linear features. The existing technique called Convolutional Neural Network (CNN) [38] is designed for a smart healthcare system for monitoring the MFH. Multiple features extraction techniques are used to calculate both non-linear and non-linear features. The CNN network consists of six CLs, max-pooling, single flatten layer and output layer for the prediction process. The work [38] uses homogenous i.e. having the same neuron type in the entire network and based solely on the linear-neuron model and didn't focus on non-linear features for better prediction. The data collected from 45 patients before labor and 55 patients during labor and validated the performance of CNN with existing ML techniques. But the results are not satisfactory, while using one-dimensional CNN, which is advanced network called DC-GAN is incorporated with a smart healthcare system for monitoring the MFH. In addition, this proposed model uses the standard datasets for the prediction process and compared them with existing techniques such as CNN [38], Long-Short term memory (LSTM), etc.

1) *Validation analysis of the proposed model without feature extraction:* Initially, the experimental results of DC-GAN are compared with CNN, LSTM, RNN, and DNN in terms of sensitivity, specificity, and precision, which are shown in Table VII and Fig. 6.

The proposed DCGAN model achieved 62.32% of precision and 77% of sensitivity and specificity. But, the DNN and LSTM model achieved nearly 46% to 50% of sensitivity, specificity, and precision. The reason is that the LSTM model requires more memory and takes a long time for training and it is easy to overfit. The DNN is also extremely expensive to train due to complex data models and it requires expensive GPUs. The CNN and RNN models achieved nearly 51% to 52% of sensitivity, specificity, and the precision.

TABLE VII. COMPARATIVE ANALYSIS OF VARIOUS DL ALGORITHMS WITHOUT FEATURE EXTRACTION

DL Algorithms	Parameter Metrics		
	Sensitivity	Specificity	Precision
DC-GAN	77.66	77.66	62.32
CNN	52.27	52.27	51.65
RNN	51.52	51.52	50.73
LSTM	50.87	50.52	48.56
DNN	47.85	48.65	46.5

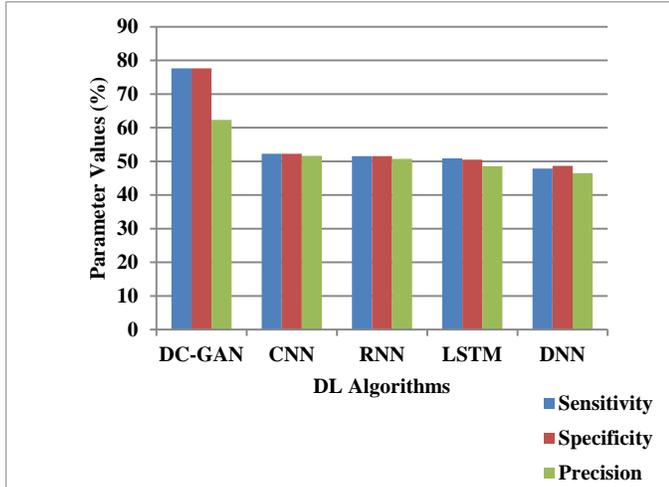


Fig. 6. Graphical Representation of Proposed DCGAN without Feature Extraction.

The CNN model requires more training data for accurate results, but now it has only three inputs of the emergency subsystem and provides low performance. The complex training procedures are presented in the RNN model and hence it achieved less than 60% of all parameters. Table VIII and Fig. 7 show the performance analysis of DCGAN and existing DL techniques in terms of accuracy and F1-score by implementing without feature extraction techniques.

In the accuracy analysis, DNN and LSTM achieved less performance (only 74%) than other DL techniques, and proposed DCGAN achieved high performance (i.e. 89.03%) than CNN and RNN techniques. But, the same proposed method achieved very low performance in terms of F1-score (i.e. 65.92%), because the three inputs of the emergency-subsystem are directly given to the DCGAN model and it provides low performance since it is a one-dimensional model. Even though the proposed model achieved less performance, it provides better performance than CNN, RNN, LSTM, and DNN. The reason is that these existing techniques are unable to handle the inputs of the emergency subsystem due to its structures. DNN and LSTM model achieved nearly 46% to 49% of F1-score, whereas RNN and CNN achieved nearly 50% of F1-score.

2) *Validation analysis of the proposed model with feature extraction:* In this experimental analysis, the importance of feature extraction is revealed by implementing it with all the existing DL and proposed DCGAN models in terms of various parameters. Table IX and Fig. 8 show the performance

analysis of different DL techniques in terms of specificity, precision, and sensitivity.

TABLE VIII. COMPARATIVE ANALYSIS OF VARIOUS DL ALGORITHMS WITHOUT FEATURE EXTRACTION

DL Algorithms	Accuracy	F1-Score
DC-GAN	89.03	65.92
CNN	81.66	51.72
RNN	83.63	50.24
LSTM	75.95	49.85
DNN	73.54	46.21

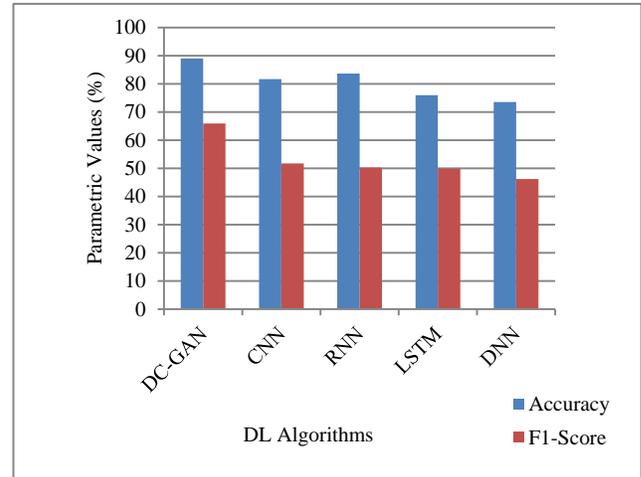


Fig. 7. Graphical Representation of Proposed DCGAN in terms of Accuracy and F1-score by considering without Feature Extraction.

TABLE IX. COMPARATIVE ANALYSIS OF VARIOUS DL ALGORITHMS WITH FEATURE EXTRACTION

DL Algorithms	Sensitivity	Specificity	Precision
DC-GAN	98.27	96.70	97
CNN	95.95	92.63	93
RNN	97.11	94.62	95
LSTM	91.62	84.57	85
DNN	90.87	81.23	83

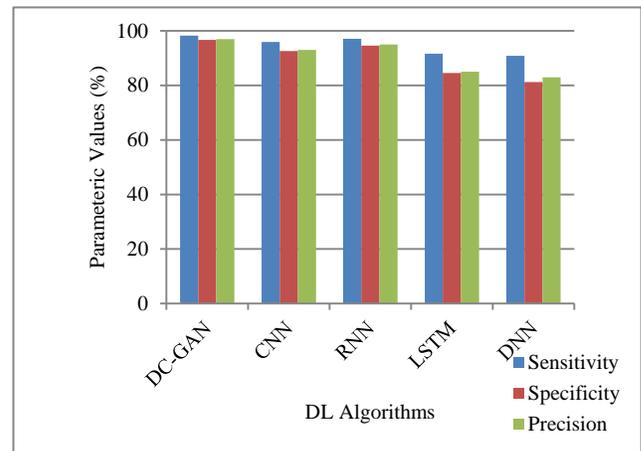


Fig. 8. Graphical Representation of Proposed DCGAN with Feature Extraction.

When all the DL techniques are implemented with feature extraction, its performance is highly improved in terms of precision, specificity, and precision, while comparing it without feature extractions. This shows that feature extraction plays a major role in this study and the one-dimensional CNN, DCGAN, RNN, and DNN are used here and it requires feature extraction techniques for better performance. In the analysis of sensitivity, every DL technique achieved more than 90% and the proposed DCGAN achieved 98.27%. The specificity of DCGAN is less (i.e. 96.70%) than the precision (i.e.97%) of proposed DCGAN, but its performance is highly improved while compared with existing DL techniques such as CNN, RNN, LSTM, and DNN. The CNN and RNN achieved nearly 92% to 95% of specificity and precision, whereas LSTM and DNN achieved nearly 81% to 85% of specificity and precision. This analysis shows that the proposed DCGAN model achieved better performance than other existing techniques. Table X and Fig. 9 show the experimental analysis of various DL algorithms in terms of accuracy and F1-score by considering feature extraction techniques.

TABLE X. COMPARATIVE ANALYSIS OF VARIOUS DL ALGORITHMS WITH FEATURE EXTRACTION

DL Algorithms	Accuracy	F1-Score
DC-GAN	97.50	96
CNN	96.03	94
RNN	96.78	95
LSTM	89.98	86
DNN	88.58	84

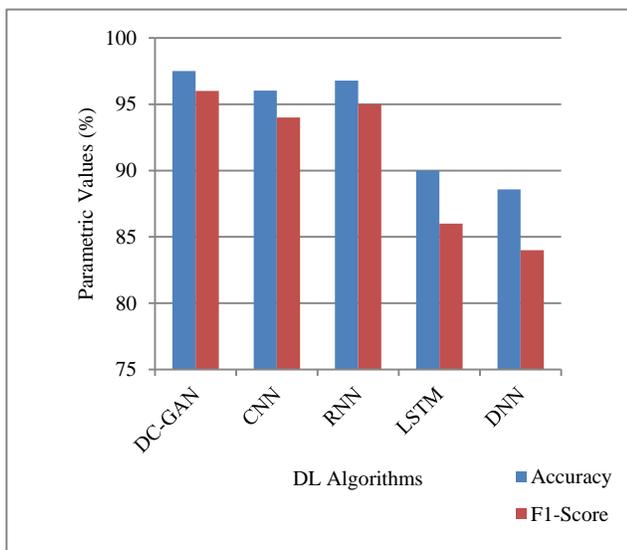


Fig. 9. Graphical Representation of Proposed DCGAN in Terms of Accuracy and F1-score by considering with Feature Extraction.

In the analysis of the F1-score, the LSTM and DNN achieved nearly 85%, whereas CNN and RNN achieved nearly 94% to 95%, but the proposed DCGAN achieved 96%. The reason for the better performance of the proposed model is to introduce the conditional-based network in the DC-GAN, where the other techniques use only its basic architecture. In the experiments of accuracy, the DNN and LSTM techniques

achieved 89%, the CNN and RNN achieved nearly 96%, and the proposed DCGAN network achieved 97.50% of accuracy while implementing feature extraction. Fig. 10 shows the comparative analysis of various DL algorithms in terms of accuracy by considering with and without feature extraction techniques for better validation purposes.

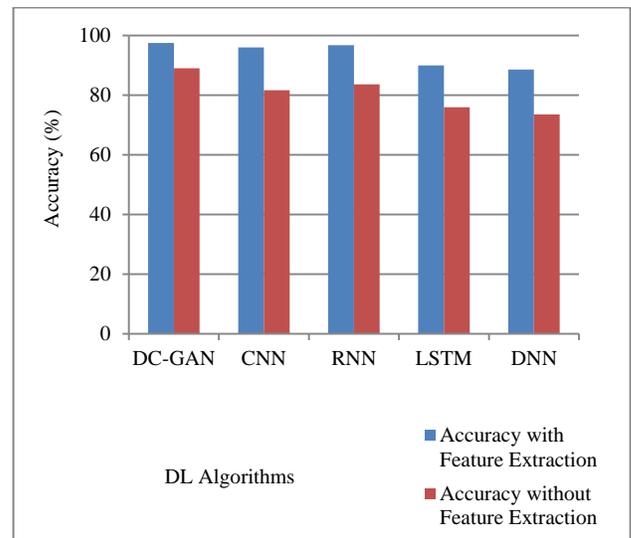


Fig. 10. Comparative Analysis of DL Techniques with and without Feature Extraction Techniques in Terms of Accuracy.

V. CONCLUSION

According to the diagnostic system's fog computing framework, an integrated solution is provided for high-risk monitoring of maternal patients, where this process relies on IoT-network-based sensors. A large data-analysis module for feature extraction techniques are used to extract features from the observed signals, and finally, an in-depth DCGAN-based classification system for predicting maternal-fetal health by the possible outcomes as PC1 to PC6 for performance analysis. The main contribution of this work in the Emergency Diagnostics Branch is the design of a set of hypothetical rules that can achieve more than 80% of the area under the curve for fetal, maternal, and emergencies. A total of 6 possible outcomes from the prediction model were considered. Compared with existing DL techniques such as CNN, LSDM, RNN, and DNN, the results show that DCGAN techniques and homogeneous systemic extraction performed better in the maternal and fetal and stage II. In the future, the efficiency of the emergency diagnostic system can be improved by modifying the selected model using different hypothetical rules.

REFERENCES

- [1] WHO. Maternal mortality, Retrieved on 2019. <http://www.who.int/mediacentre/factsheets/fs348/en/>.
- [2] WHO. Aging and health, Retrieved on 2019. <http://www.who.int/mediacentre/factsheets/fs404/en/>.
- [3] L. Atzori, et al., The internet of things: a survey, Computer Network. 54 (15) (2010) 2787–2805.
- [4] J.A. Stankovic, et al., Research directions for the internet of things, IEEE Internet Things J. 1 (1) (2014) 3–9.

- [5] J. Gubbi, et al., Internet of things (IoT): A vision, architectural elements, and future directions, *Future Generation Computer System*. 29 (7) (2013) 1645–1660.
- [6] R. Mieronkoski, et al., The internet of things for basic nursing care—a scoping review, *International Journal of Nursing Studies*. 69 (2017) 78–90.
- [7] W. Gyselaers, V. Storms, L. Grieten, New technologies to reduce medicalization of prenatal care: a contradiction with realistic perspectives, *Expert review of medical devices*. 13 (8) (2016) 697–699.
- [8] Y. Lu, X. Zhang, X. Fu, F. Chen, K.K.L. Wong, Ensemble machine learning for estimating fetal weight at varying gestational age, in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, 2019*, pp. 9522–9527.
- [9] J. Li, L. Huang, Z. Shen, Y. Zhang, M. Fang, B. Li, X. Fu, Q. Zhao, H. Wang, Automatic classification of fetal heart rate based on convolutional neural network, *IEEE Internet Things J.* 6 (2) (2019) 1394–1401.
- [10] Y. Lu, X. Zhang, L. Jing, X. Li, X. Fu, Estimation of the fetal heart rate baseline based on singular spectrum analysis and empirical mode decomposition, *Future Generation Computer System* 112 (2020) 126–135.
- [11] Y. Lu, Y. Gao, Y. Xie, S. He, Computerised interpretation systems for cardiotocography for both home and hospital use, in: *Proceedings of the 31st IEEE International Symposium on Computer-Based Medical Systems, CBMS 2018, IEEE, 2018*, pp. 422–427.
- [12] Y. Lu, Y. Qi, X. Fu, A framework for intelligent analysis of digital cardiotocographic signals from IoMT-based fetal monitoring, *Future Gener. Comput. Syst.* 101 (2019) 1130–1141.
- [13] Lu, Y., Fu, X., Chen, F., & Wong, K.K., “Prediction of fetal weight at varying gestational age in the absence of ultrasound examination using ensemble learning”, *Artificial intelligence in medicine*, 102, 101748 . (2020).
- [14] C.Perera, C.H. Liu, S. Jayawardena, The emerging internet of things marketplace from an industrial perspective: A survey, *IEEE Trans. Emerg. Top. Comput.* 3 (4) (2015) 585–598.
- [15] H. Ghasemzadeh, E. Guenterberg, R. Jafari, Energy-efficient information-driven coverage for physical movement monitoring in body sensor networks, *IEEE Journal of. Selective. Areas Communication*. 27 (1) (2009) 58–69.
- [16] K. Wac, M.S. Bargh, B. Jan F. Van Beijnum, R.G. Bulst, P. Pawar, A. Peddemors, Power- and delay-awareness of health tele-monitoring services: the mobi-health system case study, *IEEE Journal on Selected Areas in Communications*. 27 (4) (2009) 525–536.
- [17] G. Chiarini, P. Ray, S. Akter, C. Masella, A. Ganz, “mHealth technologies for chronic diseases and elders: A systematic review”, *IEEE Journal on Selected Areas in Communications*. 31 (9) (2013) 6–18.
- [18] Oh, J. Cha, M. Ji, H. Kang, S. Kim, E. Heo, J.S. Han, H. Kang, H. Chae, H. Hwang, S. Yoo, Architecture design of healthcare software-as-a-service platform for cloud-based clinical decision support service, *Healthcare Inf. Res.* 21 (2) (2015) 102–110.
- [19] A.J. Jara, M.A. Zamora-Izquierdo, A.F.S. and, Interconnection framework for mhealth and remote monitoring based on the internet of things, *IEEE Journal on Selected Areas in Communications*. 31 (9) (2013) 47–65.
- [20] M. Barcelo, A. Correa, J. Llorca, A.M. Tulino, J.L. Vicario, A. Morell, IoT-Cloud service optimization in next-generation smart environments, *IEEE Journal on Selected Areas in Communications*. 34 (12) (2016) 4077–4090.
- [21] Y. Huang, M.Y. Hsieh, H.C. Chao, S.H. Hung, J.H. Park, “Pervasive, secure access to a hierarchical sensor-based healthcare monitoring architecture in wireless heterogeneous networks”, *IEEE Journal on Selected Areas in Communications*. 27 (4) (2009) 400–411.
- [22] D. Niyato, E. Hossain, S. Camorlinga, Remote patient monitoring service using heterogeneous wireless access networks: architecture and optimization, *IEEE Journal on Selected Areas in Communications*. 27 (4) (2009) 412–423.
- [23] M.A. Salahuddin, A. Al-Fuqaha, M. Guizani, K. Shuaib, F. Sallabi, Softwarization of internet of things infrastructure for secure and smart healthcare, *Computer* 50 (7) (2017) 74–79.
- [24] Allahem, Hisham and Srinivas Sampalli, “Automated uterine contractions pattern detection framework to monitor pregnant women with a high risk of premature labor”, *Informatics in Medicine Unlocked*, vol.20, pp.100404, 2020.
- [25] Liu L, Jiao Y, Li X, Ouyang Y, and Shi D, “Machine learning algorithms to predict early pregnancy loss after in vitro fertilization-embryo transfer with fetal heart rate as a strong predictor”, *Computer Methods and Programs in Biomedicine*, vol.196, no.105624, 2020.
- [26] Testa, A. and Jackson, D.B., 2020. Incarceration exposure and maternal food insecurity during pregnancy: Findings from the Pregnancy Risk Assessment Monitoring System (PRAMS), 2004–2015. *Maternal and child health journal*, 24(1), pp.54-61.
- [27] Azimi I, Pahikkala T, Rahmani A. M, Niela-Vilén H, Axelin A, and Liljeberg P, “Missing data resilient decisionmaking for healthcare IoT through personalization: A case study on maternal health”, *Future Generation Computer Systems*, vol.96, pp.297-308, 2019.
- [28] Mourad M, A. Diab, M. Khalil, and C. Marque, “Pregnancy/Labor Discrimination and Monitoring: An Investigation Using Nonlinear Methods”, *International Arab Conference on Information Technology (ACIT)*, Werdanye, Lebanon, pp.1-4, 2018.
- [29] J. A. Lobo Marques, P. C. Cortez, J. P. D. V. Madeiro, S. J. Fong, F. S. Schlindwein, and V. H. C. D. Albuquerque, “Automatic Cardiotocography Diagnostic System Based on Hilbert Transform and Adaptive Threshold Technique,” in *IEEE Access*, vol. 7, pp. 73085-73094, 2019.
- [30] Ayres-de-Campos, D., Spong, C.Y., &Chandrahara, E. (2015). “FIGO consensus guidelines on intrapartum fetal monitoring: Cardiotocography”, *International Journal of Gynecology& Obstetrics*, 131.2015.
- [31] Q. Zhang, L. T. Yang, Z. Chen, P. Li and F. Bu, “An Adaptive Dropout Deep Computation Model for Industrial IoT Big Data Learning With Crowdsourcing to Cloud Computing,” in *IEEE Transactions on Industrial Informatics*, vol. 15, no. 4, pp. 2330-2337, April 2019.
- [32] Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv 2016*, arXiv:1511.06434.
- [33] Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv 2014*, arXiv:1411.1784.
- [34] Sugawara, Y., Shiota, S., & Kiya, H. “Convolutional Neural Networks Without Any Checkerboard Artifacts”,. 2018 26th European Signal Processing Conference (EUSIPCO), 1317-1321. 2018.
- [35] Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*; Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q., Eds.; Morgan Kaufmann Publishers Inc.: Burlington, MA, USA, 2014.
- [36] A. Asuncion and D. J. Newman. “UCI machine learning repository,”[<http://www.ics.uci.edu/~mlern/MLReposit-ory.html>] University of California, School of Information and Computer Science, Irvine, CA, 2007.
- [37] Ayresde, C.D., Bernardes, J., Garrido, A., et al.: SisPorto 2.0: “A program for automated analysis of cardiotocograph”,*The Journal of Maternal-Fetal& Neonatal Medicine*. 9(5), 311–318 (2000).
- [38] Marques, J.A.L., Han, T., Wu, W., do Vale Madeiro, J.P., Neto, A.V.L., Gravina, R., Fortino, G. and de Albuquerque, V.H.C., 2020. IoT-based Smart Health System for Ambulatory Maternal and Fetal Monitoring. *IEEE Internet of Things Journal*,.2020.

Periapical Radiograph Texture Features for Osteoporosis Detection using Deep Convolutional Neural Network

Khasnur Hidjah¹

Department of Computer Science and Electronics
Universitas Gadjah Mada, Yogyakarta, Indonesia
Departement of Computer Science
Universitas Bumigora, Mataram, Indonesia

Agus Harjoko^{2*}

Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia

Moh. Edi Wibowo³

Department of Computer Science and Electronics
Universitas Gadjah Mada
Yogyakarta, Indonesia

Rurie Ratna Shantiningsih⁴

Departement of Dentomaxillofacial Radiology
Universitas Gadjah Mada
Yogyakarta, Indonesia

Abstract—Currently, research for osteoporosis examination using dental radiographic images is increasing rapidly. Many researchers have used various methods from subject data. It indicates that osteoporosis has become a widespread disease that should be studied more deeply. This study proposes a deep Convolutional Neural Network architecture as a texture feature of dental periapical radiograph for osteoporosis detection. The subject of this study is postmenopausal Javanese women aged over 40 and data measurement result of Bone Mineral Density. The proposed model is divided into stages: 1) stage image acquisition and RoI selection, 2) stage feature extraction and classification. Various experiments with the number of convolution layers (3 layers to 6 layers) and various input block sizes and other hyper parameters were used to get the best model. The best model is obtained when the input image size is greater than 100 and less than 150 and a five of convolution layer, as well as other hyper parameters, including epochs=100, dropout=0.5, learning rate=0.0001, batch size= 16 and loss function using Adam's optimization. Validation and testing accuracy achieved by the best model is 98.10%, and 92.50. The research shows that the bigger images provide additional information about trabecular patterns in normal, osteopenia and osteoporosis classes, so that the proposed method using deep convolutional neural network as textural feature of the periapical radiograph achieves a good performance for detection osteoporosis.

Keywords—Osteoporosis; dental periapical radiograph; convolutional neural network; texture features; bone mineral density

I. INTRODUCTION

Osteoporosis is defined as a systemic skeletal disease characterized by low bone mass and micro-architectural deterioration of bone tissue [1]. Therefore, osteoporosis will increase the consequences of bone fragility and susceptibility to fracture, especially for those over 50 years of age. Once a fracture happens to someone with osteoporosis, life will be greatly affected due to disability to move and prolonged

healing process. Finally, this reduces a person's quality of life and causes various economic and social problems [2].

For example, if the injured person works as a driver or as a labor worker, he might have to retire and find some disk-related job that is not easy to obtain. In some other cases, the injured person might become severely disabled and require continuous assistance, which might burden his family. Therefore, preventative measures and early treatment of osteoporosis [3-4] are the best options to address these issues. Practical scientific and technological methods to support osteoporosis diagnosis, in this context, will provide much help to overcome the disease and reduce its negative impacts.

The most accurate BMD examination and made the gold standard by World Health Organization (WHO) is using Dual Energy X-Ray Absorptiometry (DEXA). However, access to this method is still limited in many countries. BMD examination is often available in central hospitals only, and its cost is often too expensive for many people in rural areas. Furthermore, BMD is not able to reveal the internal structure of fractured bones [4-5]. Researchers, therefore, have attempted to develop alternative methods that are more practical and more widely accessible. Several studies have found that dental data demonstrate a high correlation with BMD measurements [6-16]. The data include panoramic and periapical radiographs. Besides that, of the use widespread of periapical radiographs in dental care for the elderly with increased life expectancy and the number of studies according to BMD estimates and screening for osteoporosis using periapical radiographs. It is expected to provide benefits, namely the architecture that has been produced can then be used as an architectural model in pre-train medical images for different cases specifically using medical images with the same characteristics, which tend to be low resolution. In addition, it can also help patients who perform dental checkup at the dentist to be given a referral to chiropractors for further

*Corresponding Author

checkup, so that it can be detected early if you have osteoporosis.

Periapical radiograph is a dental radiograph technique that can image four to five teeth and their respective areas on one intraoral X-ray films [7]. At a microstructure level, images of trabecular jaw bones often show visual patterns closely related to the general condition of other bones [8,10]. Dental data have therefore become promising sources to predict Bone Mineral Density (BMD) measurements accurately. Periapical radiographs, in particular, will become the focus of this research since these radiographs are much more affordable and generally are more available. Fig. 1 shows an example of a periapical radiograph of the mandible.

The remaining part of this research is organized as follows. Section II explains some related work on osteoporosis detection using dental periapical radiograph, while Section III provides a more detailed description of the proposed method and training CNN's. Section IV contains result and discussion of the proposed method, and Section V concludes the research.

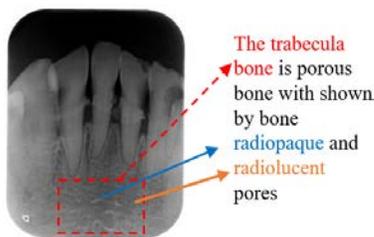


Fig. 1. Periapical Radiograph Image of the Mandibl.

II. RELATED WORK

Several studies have been carried out to examine osteoporosis using periapical radiographs. In [8] has made one of attempts to predict bone mass from such radiographs. Data and the corresponding labels (BMD measurements) are collected from 60 postmenopausal women aged over 40. This research extracts and combines visual features such as areas, lengths, and peripheries of “bright” blobs, numbers of terminal/branch points, and clinical data such as ages, body sizes, and calcium intakes. The overall number of pixels in the RoI divides that of black pixels results in the visual features of the trabecular area. The total amount of pixels in the external of the trabeculae in the binary images displays the periphery. In other words, it represents a proportion of the entire area of the trabeculae or the entire area of RoI. The skeletonized image computes the complete area of the skeletonized trabeculae (total amount of black pixels), the number of terminal points (free ends, that is, black pixels with only one neighboring black pixel), and the number of branch points (crossing points, that is, black pixels with three or more neighboring black pixels). These are used to represent the percentage of trabecular length, area and perimeter. Classification and regression tree analysis (CART) uses patients in groups of normal or low bone mass categories. It is evident from the CART analysis of clinical and radiographic features that the main element to categorize patients as having normal or low bone mass were age (± 42.5 years) and the number of terminal points as a function of the periphery (± 0.09). This algorithm conscientiously

distinguished 22 normal patients by BMD (specificity = 100%) and 31 patients with low bone mass (sensitivity = 81%). The total accuracy was 88.33%. A denomination of the corresponding predicted and actual bone category, the weighted kappa index, was 0.76. To identify women with low BMD, trabecular morphology analysis was an alternative. Another research [9] was also conducted displaying the combination of upper and lower jaw radiographs from 505 postmenopausal women aged 45–70 years. Dense, sparse or mixed trabecular patterns were identified by five observers. The gradings were integrated into a single averaged observer score per jaw in which the RoI can be identified on each by scanned radiographs. The RoIs compounded with image analysis software measured 25 photographs' characteristics. Pearson correlation and multiple linear regressions, which were used in identifying the averaged observer score, showed that 14 image features were significantly correlated with the observer judgment for the two jaws. Other features, which give details of osteoporotic patients with fewer but bigger marrow spaces than controls, are less compatible with the sparseness of the trabecular pattern than a rather crude measure for a structure such as the average grey value. To sum up, the human concept of sparseness is emanated more from average grey values of the RoI than from geometric details within the RoI. In [11], the bone mass prediction on porosities, connectivity, and orientations of porous was shown in trabecular images and a combination of the anthropometric features (weight, height, age, body mass index). While a decision tree was used to select the feature, a backpropagation artificial neural network was used for classification. By combining age, weight, height, body mass index and features of trabecular morphology interdental bone, identifying postmenopausal women with low bone mass are much easier. In this study, however, age is considered one of the biggest contributors to loss of bone mass. Porosity, the oblique porous, and the vertical porous are crucial porous features. This study distinguishes anthropometric and radiographic features, which then is analyzed individually. Both anthropometric features and the radiographic features have high accuracy with 80.33% and 87.04%, respectively. This work has been extended further [12] that combine data from periapical and panoramic radiographs. Furthermore in [13], a method for osteoporosis identification based on the validated trabecular area was presented on digital dental radiographic images. The image RoI of the validated trabecular area on the images should be obtained through a sequence of morphological operations, which is then evaluated using the Dice similarity method. In analyzing osteoporosis, a mineral density is estimated using dual X-ray absorptiometry in two areas and by extracting RoI through statistical features (deviation, entropy, homogeneity, and correlation). Feature extraction and feature selection are used to analyze the four features. The selection process applies the C4.5 feature selection method. Subsequently, to estimate the existence of osteoporosis, a multilayer perceptron of statistical texture analysis is employed. 0.8924. is obtained as the result of the average dice similarity coefficient for all of RoIs. The most suitable method in this proposed study, achieving an accuracy of 87.87%, is a multilayer perceptron classifier.

A study on the analysis of the mandibular trabecular structure in postmenopausal women using periapical

radiographs was conducted by [14]. The mandibular trabecular structure parameter used was the thickness of the trabeculae compared with the results of BMD DXA measurements in the lumbar (spine) and femoral (hip) areas. Determine the RoI manually with a size of 100pixels x 100pixels with a position 2mm from the apical edge of the left and right posterior parts of the lower jaw (mandible). Measurement of trabecular thickness using slice geometry features, namely bone area fraction size by dividing the number of pixels classified as bone by the total pixel area in RoI and trabecular thickness 2D is the average trabecular width in RoI. These two parameters were correlated with femoral and lumbar BMD values. Then the measurement results with statistical analysis showed a statistically significant difference between the normal group and the osteoporosis group compared to the normal group and the osteopenia group, so it can be said that thinning of the trabecular structure is more clearly seen in postmenopausal women with osteoporosis, with bone quality that can be detected earlier using the trabecular thickness parameter.

Although several methods have been proposed to examine osteoporosis using dental periapical radiographs, the methods generally rely on morphological analysis and geometric features of the images [8–14]. Only a little work has been conducted to investigate the effectiveness of texture features such as [15–16] that employed features derived from the Gray Level Co-occurrence Matrix (GLCM). However, the employed features are considered to be handcrafted, which might be suboptimal for the given problem. Therefore, the facts mentioned above have suggested further investigation on the use of texture features, particularly those that are directly learned from data. This research proposes deep learning in the analysis of texture features for the prediction of osteoporosis. Deep learning has worked effectively in many areas, including computer vision, hyperspectral image processing, medical image analysis [17], and natural language processing include in tuning for hyperparameters online [18]. Compared to conventional methods such as support vector regressors and multi-layer perceptron, based on feature, deep learning has some advantages, such as working on two-dimensional data directly, less susceptibility to local optimal, and the ability to learn texture features from data [19]. The other advantages of DCNN are transferability connections and sparse connections. The transferability connections are certain layers of network architecture that can reproduce weights for different tasks. Sparse connections are infrequent connections that can reduce redundant connectivity, thereby reducing computing costs [21].

One particular model of a deep learning is convolutional neural network or also called deep convolutional neural network (DCNN). Since 2012, DCNN [17–20] has been led to a series of breakthroughs for image classification [22]. Deep learning-based computer-aided diagnosis for breast cancer [23] and lung cancer [24] has been applied in radiology. In addition, there are many other studies that use DCNN to detect or classify diseases, such as [25] comparing the performance of three CNN models (models VGG19, Resnet50v2 and Densenet201) with X-rays data sets of patients with COVID-19, pneumonia, and tuberculosis with a large number of data sets.

By considering several capabilities and advantages of DCNN, our contributions are:

- We designed and determined the best DCNN configuration model to extract in-depth features from dental periapical radiograph images from multiple image block sizes and multiple convolution layers with varying hyper parameter.
- The result of DCNN architecture or configuration can be used to detect osteoporosis disease.

Measured the effectiveness of the best model by conducting trials using data from previous researchers to achieve state-of-the-art performance for the detection or classification of osteoporosis using dental periapical radiograph.

III. METHOD

This study proposes a deep Convolutional Neural Network architecture as a texture feature of dental periapical radiographs that can be used for osteoporosis detection. An extensive examination of the network is conducted to obtain the optimal network configuration and hyperparameters, which include input image size, number of kernels, filter kernel size, dropout value, and learning rate value. The system results will be compared with results of BMD measurements in the femur and lumbar areas using DXA.

The proposed model is broadly divided into two stages, namely the training stage and the testing stage. Each stage consists of several processes, namely, image acquisition and ROI selection, feature extraction and classification (see Fig. 2).

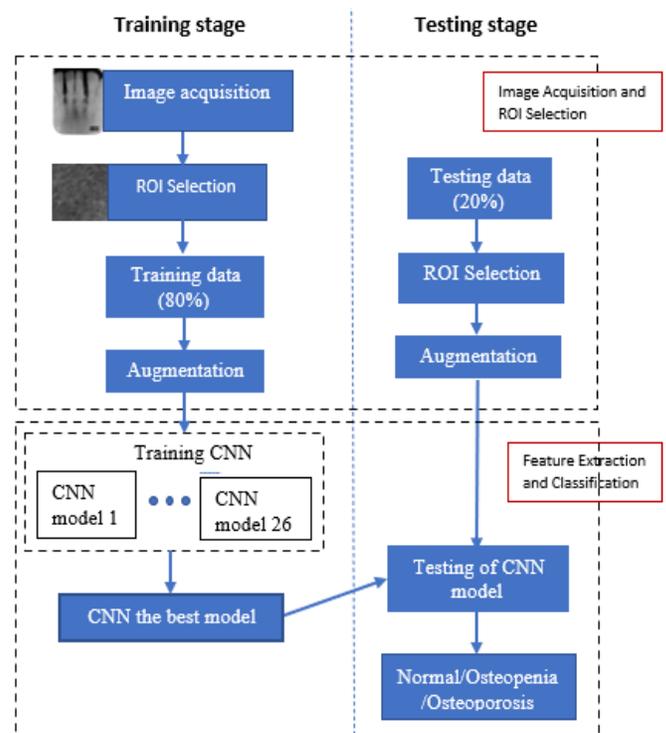


Fig. 2. Proposed Method for Osteoporosis Detection.

A. Image Acquisition and RoIs Selection

We used a dental X-ray device to obtain digital periapical radiographs of mandibular anterior teeth of 31 postmenopausal women aged over 40 years old. This research makes use of the Villa Sistemi device with an electrical specification of 70 kVp/8 mA that uses photostimulable phosphorus plates (PSP) as image receptors. All periapical radiographs were processed from the Radiology Department, Prof. Soedomo Dental Hospital of Universitas Gadjah Mada using the DBSWin4.5 (Dürr Dental, Bietigheim-Bissingen) to produce digital grayscale images in the JPG format. All images are assessed for quality assurance by a dentist (Fig.3).

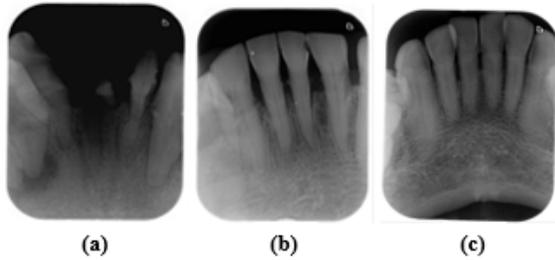


Fig. 3. Periapical Radiograph of the Mandible (a) Normal, (b) Osteopenia, and (c) Osteoporosis.

Regions of interests (ROIs) are then selected semi-manually from the images to obtain the most appropriate parts for further processing. The selection procedure marks the upper left corner of the trabecular area then moves to the lower right to form the maximal square that can be extracted from the images. so that ROIs is obtained with various sizes, at least 250x250 pixels Assuming that trabecular areas' sizes do not vary significantly across people, all the extracted ROIs are resized to a standard size of 250 x 250 pixels. It can be considered as a normalization step favoured by subsequent processes. Fig. 4 shows the ROIs selection process as well as the resized images. The resized ROIs are divided further into overlapping blocks (with 10 pixels overlap), each of which will become an input to a convolutional neural network. This process is called augmentation.

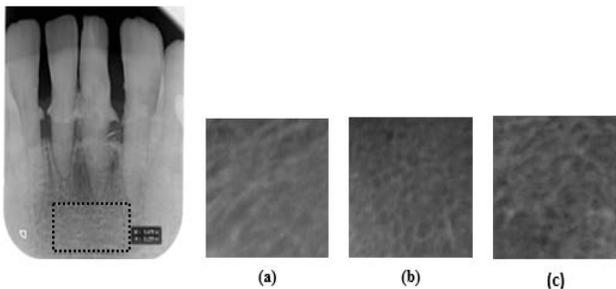


Fig. 4. ROIs Selection and Resizing (a) Normal, (b) Osteopenia and (c) Osteoporosis.

B. Features Extraction and Classification

Feature extraction and classification is performed using a deep convolutional neural network (CNN) which takes image blocks as input and produces a prediction class as output. The prediction class in this study consisted of normal (N), osteopenia (Oa), and osteoporosis (Os) which was a further

decrease in the bone mass of the examined subjects. The deep CNN configuration used is shown in Fig. 5 or details can be seen in Table I. The deep CNN configuration consists of five building blocks, namely convolution layer, activation layer, pooling layer, fully connected layer, and soft-max layer. The first convolution layer uses a kernel size of 5x5, while the second to fifth convolution layers use a 3x3 kernel size. While in the pooling layer, all layers use a 2x2 kernel.

The Deep CNN configuration showed at Fig. 5 and Table I is the best model from the results of experiments that have been carried out on each block size and convolution layer size, activation layer, pooling layer including number of the kernel used (6, 16, 32, 64, 128) and the filter kernel size for each convolution layer as well as several parameters such as learning rate value, dropout value and number of epochs (as shown in Fig. 6-9 and Table II-III).

CNN's themselves are inspired by a neuro-biological process in which connectivity patterns between neurons resemble the visual cortex model [26] and [27]. CNN's work on two-dimensional data of multiple depths and operate in a layer-by-layer order [28].

Convolution layers serve to extract features from an input image (edges, corners, or crosses) using responses to some special character presenting in the input. Activation layers determine "relevant" convolutional kernels. The layers produce stacks of feature maps, each of which parts within the produced feature maps to be used in subsequent processing. Relevant parts, in this case, will be "active" after passing through the activation function, which is the rectified linear units (ReLU).

The ReLU layer is an activation function obtained through the equation.

$$ReLU(x) = \max(0, x) \tag{1}$$

Where x is the input to the neuron and the transfer function is finely approximated to the rectifier into an analytic function.

Pooling layers help the network avoid overfitting by reducing some network parameters and the respective computations. The pooling layers work as a non-linear down-sampling process that divides outputs of activation layers into subregions and collects maximum values from the subregions. From an $n \times n$ input, with n represented of size of image and k represented from size of kernel, a pooling layer will produce an $\left(\frac{n}{k}\right) \times \left(\frac{n}{k}\right)$ output.

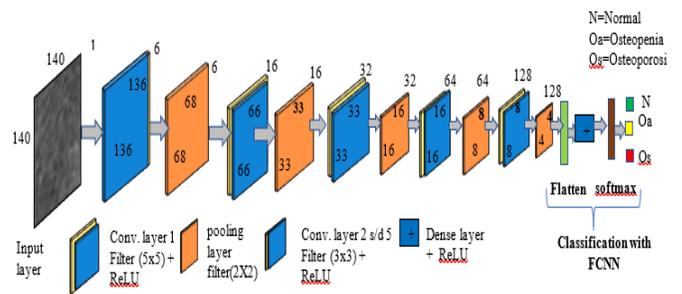


TABLE I. CONFIGURATION MODEL OF CNN FOR OSTEOPOROSIS DETECTION

LAYER TYPE		INPUT SIZE	NUMBER OF KERNEL	FILTER KERNEL SIZE	OUTPUT SHAPE
Input layer	input	140x140x1	-		140x140x1
Conv. layer 1	convolution	140x140x1	6	5x5	136x136x6
	activation	136x136x6	6		136x136x6
	max pooling	136x136x6	6	2x2	68x68x6
Conv. layer 2	convolution	68x68x6	16	3x3	66x66x16
	activation	66x66x16	16		66x66x16
	max pooling	66x66x16	16	2x2	33x33x16
Conv. layer 3	convolution	33x33x16	32	3x3	33x33x32
	activation	33x33x32	32		33x33x32
	max pooling	33x33x32	32	2x2	16x16x32
Conv. layer 4	convolution	16x16x32	64	3x3	16x16x64
	activation	16x16x64	64		16x16x64
	max pooling	16x16x64	64	2x2	8x8x64
Conv. layer 5	convolution	8x8x64	128	3x3	8x8x128
	activation	8x8x128	128		8x8x128
	max pooling	8x8x128	128	2x2	4x4x128
	Dropout 0.5	4x4x128			4x4x128
FCNN	flatten	4x4x128			2048
	dense128	128			128
	activation (ReLU)	128			128
	Dropout 0.5	128			128
	dense3	3			3
output softmax	activation (softmax)	3			3

In CNN's, a convolution layer is normally tied with an activation layer and a pooling layer (showed Fig. 5). This bundle is repeated several times to produce a "thick" stack of down-sampled feature maps at the end of the sequence. Fully connected layers take the vectorized (flattened) form of this

stack and are also tied with some activation layers to produce output vectors. The lengths of these vectors are normally equal to the number of prediction classes. Values within these vectors are converted into probability values by SoftMax layers at the end of CNN's. The output layer present in the last layer of CNN to the normalized exponential function or softmax is a generalization of the logical function of a k-dimensional z vector into a k-dimensional $\sigma(z)$ vector with a real number value between [0, 1]. The SoftMax function is written in the following equation (2):

$$\sigma: R^k \rightarrow [0,1]^K \quad (2)$$

$$\sigma(Z) = \frac{e^{z_j}}{\sum_{k=1}^K e^{z_k}} \text{ for } j=1, \dots, K$$

where σ is softmax notation symbol, z is a vector of the inputs to the output layer, K is dimensions of vector z, and j is the index of the output unit. Table I shows the specifications of the model configuration.

C. Training of CNNs

To build the proposed system, data are collected from postmenopausal Javanese women aged over 40 years. A total of 31 subjects have agreed to participate in the research and have signed informed consent. Ethical clearance has also been obtained from the ethics and advocacy unit of the Faculty of Dentistry of Universitas Gadjah Mada (UGM) with the number: 0061/KKEP/FGK-UGM/EC/2019. Some criteria are used to exclude subjects from the research. These include suffering from cancer with bone metastases, kidney failure, metabolic diseases (hyperparathyroidism, hypoparathyroidism, osteomalacia, renal osteodystrophy, and osteogenesis imperfecta), and taking drugs that affect bone metabolism.

After dental periapical radiographs are acquired from the Radiology Department, Prof. Soedomo Dental Hospital of Universitas Gadjah Mada, the women went for bone density examination at the Radiology Department of Dr Sardjito Hospital Indonesia. Periapical radiographs used in this research have gone through an assessment process to ensure their quality. As for bone density estimation, a lunar prodigy primo DEXA densitometer (GE Lunar Corporation, Madison, WI, USA) is used to scan the subjects' spine and femur regions at an exposure of 42 μ Gy for 1.27 minutes. Bone density values were converted into T-scores to determine osteoporosis, osteopenia, or normal. These categories were then used as the labels for the collected periapical radiographs. The conversion was conducted using the standard procedure specified by the World Health Organization (WHO). Based on BMD measurements of 13 subjects, three subjects were classified as normal, six subjects were classified as osteopenia, and four subjects were classified as osteoporosis.

Training is carried out by varying sizes of image blocks (as input), numbers of convolution layers, the use of dropout layers, and sizes of kernels. The max function is used in the pooling layers with sizes of kernels of 2x2 and strides of 2. Overlapping blocks are extracted from the collected images and are augmented by applying small random rotation, scaling, and vertical flip. This process produces thousands of image blocks that are further divided into a training set and

test set. The training set contains 80% of the overall data, while the test set contains the remaining 20% of the data. Table II, shows a summary (minimum, average, and maximum) of training and validation accuracy on 26 CNN models from 3 to 6 layers for each image size.

IV. EXPERIMENTS AND RESULT

A. Experiments Models

This section presents the experiments of the proposed model for osteoporosis detection. The first part of the experiment investigates the optimal configurations of CNNs. For this purpose, we evaluate different sizes of image blocks, namely 40x40, 50x50, 60x60, ..., 150x150 pixel. CNN's are built with 3 and 4 convolutional layers, and when the sizes of image blocks are greater than 100x100 pixel, the networks are also built with 5 and 6 convolutional layers. Two sizes of convolution kernels are applied during the experiments, i.e. 3x3 and 5x5. The strides of the convolution kernels are fixed to 2, and padding is used to maintain the inputs' original sizes during the convolution. Besides, in the 5th convolution layer block, a dropout layer of 0.5 is added. To facilitate classification by providing rules for removing or keeping neurons with probability values between 0 and 1, and the value of the learning rate used = 0.0001. See Tables II (A)-(D), Table III and Fig. 6 to Fig. 8 summarizes performance of CNNs with the different configurations.

TABLE II. TRAINING ACCURACY AND VALIDATION ON 26 CNN MODELS FOR EACH BLOCK SIZE

(A) THREE LAYERS

block size (pixel)	size of CNN layer					
	three layers (consists six models)					
	Minimum		Average		Maximum	
	Train	Valid	Train	Valid	Train	Valid
40x40	0.713	0.647	0.812	0.683	0.895	0.711
50x50	0.794	0.653	0.876	0.698	0.908	0.717
60x60	0.860	0.660	0.897	0.699	0.928	0.730
70x70	0.867	0.663	0.9025	0.701	0.931	0.740
80x80	0.880	0.664	0.913	0.687	0.939	0.713
90x90	0.903	0.668	0.935	0.705	0.961	0.738
100x100	0.911	0.656	0.858	0.679	0.954	0.697
110x110	0.881	0.624	0.858	0.529	0.974	0.723
120x120	0.929	0.712	0.950	0.731	0.962	0.742
130x130	0.914	0.730	0.942	0.746	0.973	0.771
140x140	0.909	0.764	0.951	0.800	0.969	0.816
150x150	0.919	0.798	0.958	0.814	0.977	0.833

(B) FOUR LAYERS

block size (pixel)	size of CNN layer					
	four layers (consists six models)					
	Minimum		Average		Maximum	
	Train	Valid	Train	Valid	Train	Valid
40x40	0.704	0.633	0.791	0.670	0.916	0.772
50x50	0.886	0.674	0.912	0.517	0.945	0.742
60x60	0.907	0.684	0.941	0.710	0.960	0.735
70x70	0.907	0.684	0.941	0.710	0.960	0.735
80x80	0.896	0.700	0.948	0.736	0.979	0.757
90x90	0.932	0.707	0.957	0.743	0.979	0.782
100x100	0.945	0.697	0.961	0.738	0.979	0.818
110x110	0.934	0.705	0.957	0.738	0.975	0.759
120x120	0.931	0.710	0.954	0.736	0.976	0.763
130x130	0.944	0.751	0.959	0.778	0.981	0.823
140x140	0.942	0.785	0.962	0.885	0.987	0.882
150x150	0.935	0.834	0.957	0.863	0.985	0.892

(C) FIVE LAYERS

block size (pixel)	Size of CNN Layer					
	five layers (consists seven models)					
	Minimum		Average		Maximum	
	Train	Valid	Train	Valid	Train	Valid
40x40	0.772	0.658	0.863	0.703	0.915	0.725
50x50	0.880	0.668	0.907	0.691	0.930	0.749
60x60	0.923	0.693	0.942	0.723	0.967	0.772
70x70	0.923	0.693	0.946	0.728	0.967	0.798
80x80	0.961	0.744	0.970	0.777	0.981	0.837
90x90	0.962	0.758	0.970	0.782	0.981	0.846
100x100	0.956	0.771	0.973	0.812	0.981	0.866
110x110	0.968	0.742	0.976	0.808	0.986	0.918
120x120	0.959	0.754	0.973	0.813	0.985	0.903
130x130	0.966	0.824	0.976	0.856	0.985	0.89
140x140	0.968	0.856	0.978	0.885	0.998	0.956
150x150	0.925	0.841	0.967	0.889	0.984	0.916

(D) SIX LAYERS

block size (pixel)	Size of CNN Layer					
	six layers (consists seven models)					
	Minimum		Average		Maximum	
	Train	Valid	Train	Valid	Train	Valid
40x40	not available					
50x50						
60x60						
70x70						
80x80						
90x90						
100x100						
110x110						
120x120						
130x130						
140x140	0.769	0.698	0.944	0.816	0.985	0.869
150x150	0.461	0.462	0.902	0.815	0.989	0.918

Based on 26 CNN models that have been trained (see Tables II(A)–(D), it is known that in the experiment with five layers and an image block size of 140x140, the minimum, average, and maximum values are higher than the number of convolution layers and other image blocks (see red text in Table IIC). This means that the configuration of five layers, the image block size of 140x140, and some hyper parameters as in table I are the best models of the 26 existing models. The best CNN model means a model that can differentiate trabecular patterns in the normal class, osteopenia, and osteoporosis.

Fig. 6-8 also shows experimental results in finding the best model.

Fig. 6 is a graph of the image blocks size against the accuracy of each convolutional layers. From Fig. 6, it is known that the larger the image blocks size (input blocks size greater than 100), the accuracy tends to increase, especially on the five-layer and four-layer CNN. Indicates that the image blocks size greater than 100 provides additional information on osteoporosis examination.

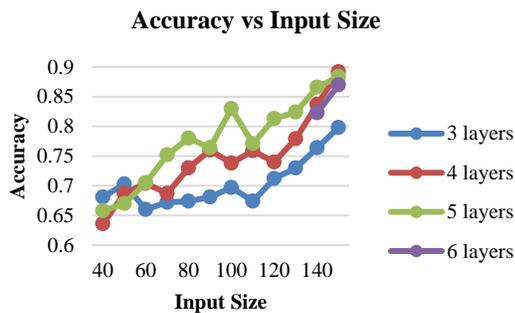


Fig. 6. Graph Accuracy Vs Input Size Graph with Kernel Size 3x3.

In Fig. 7, it is known that a CNN with five layers and an image size greater than and equal to 100 indicates increased accuracy.

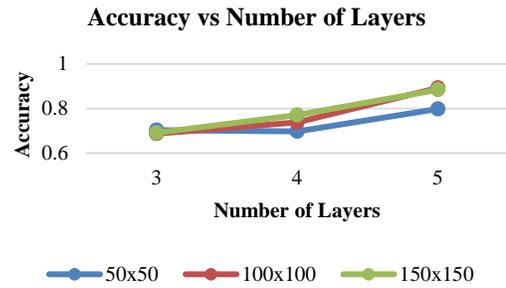


Fig. 7. Graph Accuracy Vs Number of Layers.

Fig. 8 shows the effect of the Dropout (DO) value on accuracy, the DO value = 0.5 has a better training and validation accuracy.

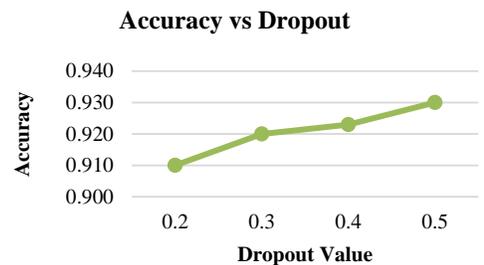


Fig. 8. Graph Accuracy Vs Dropout.

The following Table III shows the experimental results of the best model from the input image size of 40x40 pixels to 150x150 pixels with epoch = 20. And look (highlight) the best of accuracy is at 140x140 pixels.

TABLE III. TRAINING ACCURACY AND VALIDATION ON THE BEST MODELS FOR EACH BLOCK SIZE WITH EPOCH = 20

Block Size	Accuracy		Loss	
	training	validasi	Training	Validation
40x40	0.865	0.713	0.293	0.746
50x50	0.908	0.749	0.207	0.690
60x60	0.929	0.772	0.163	0.722
70x70	0.949	0.798	0.120	0.653
80x80	0.970	0.837	0.072	0.526
90x90	0.970	0.846	0.071	0.462
100x100	0.956	0.866	0.106	0.373
110x110	0.976	0.918	0.057	0.180
120x120	0.974	0.903	0.066	0.224
130x130	0.966	0.890	0.079	0.290
140x140	0.977	0.956	0.056	0.091
150x150	0.973	0.916	0.064	0.227

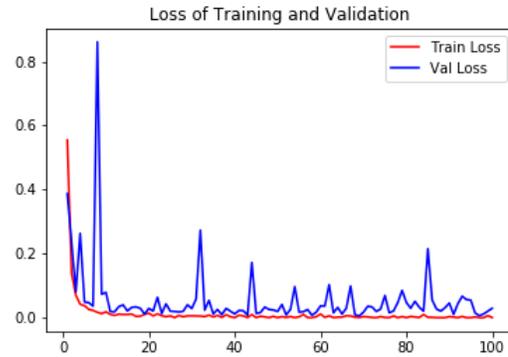
B. Training Models

The total number of training data is 40,755 images, 32,617 images for training, and 4069 images for validations. The dimension of the image input on this model is 140×140 pixels. The batch size is 16, the learning rate value is 0.0001, DO=0.5 and number of epoch = 100. Then the loss function uses Adam optimization.

Training is executed on a computer with specifications processor Intel Core i7-7500U processor specifications, 8 GB RAM, GPU: NVIDIA GeForce GTX 840, Windows 10 operating system, Python 3.7 Programming Language with an editor spyder (python3.7).

Fig. 9(a)-(b) shows the trend accuracy and loss of the training process and the validation of the best models (using architecture and configuration in Fig. 5 or Table I)

Research on osteoporosis examination using dental periapical radiograph images has been carried out with a satisfactory level of accuracy and can represent a computer-aided diagnosis system. Besides, it can enrich the extraction of textural features, which is currently known for the examination of osteoporosis with dental periapical radiograph images, which previously mostly using morphological features.



(b)

Fig. 9. (a) Graph the Accuracy of the Training and Validation Process of the Best Model, (b) Graph the Loss of the Training and Validation Process of the Best Model.

C. Testing Models

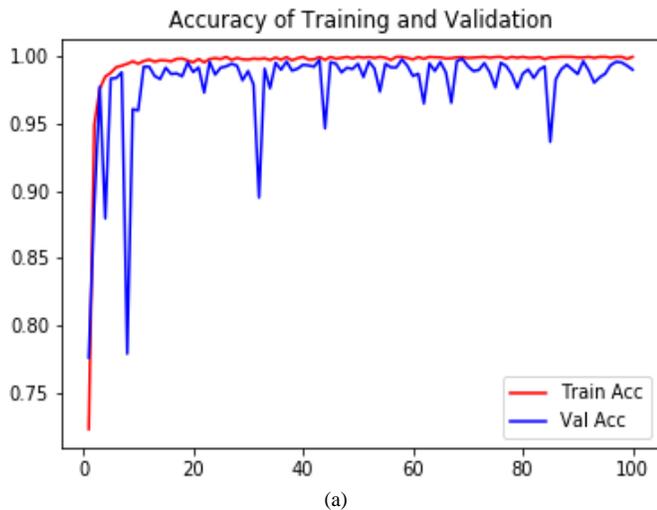
We use 4,069 images as sample test data. Model performance measured using four performance measures parameters, namely Precision, Recall, F1 score and Accuracy. The Precision, Recall, F1 score, and Accuracy values of the best model testing result see Table IV.

TABLE IV. OSTEOPOROSIS – TESTING REPORT

	Precision	Recall	F1 Score
Normal	0.88	0.88	0.88
Osteopenia	0.84	0.98	0.93
Osteoporosis	0.83	1.00	0.91
Accuracy	0.92		
Macro avg	0.97	0.97	0.97
Weighted avg	0.85	0.92	0.90

Table V shows a comparison of the performance of the osteoporosis examination between the proposed methods and those of other previous researchers. In the testing process using datasets from previous researchers [11], the dataset Augmented first so that the amount of data used in the testing process is proportional to the amount of data used in this research.

When compared to previous related work, our method has the highest validation accuracy and testing accuracy, with a validation accuracy of 98.10% and a testing accuracy of 92.50%.



(a)

TABLE V. PERFORMANCE COMPARISON OF EXAMINATION OSTEOPOROSIS

Parameters	Authors		
	<i>Licks etc. (2010) [8]</i>	<i>Sela etc. (2015) [11]</i>	<i>This Research</i>
Subject	Women over 40 years age	Women over 40 years age	Women over 40 years age
Number of training data	60 (22 normal and Osteopenia, 38 osteoporosis)	54 (11 normal, 22 Osteopenia, 21 osteoporosis)	13(3 normal, 6 osteopenia, 4 osteoporosis) augmented: 32,617 (7,527 normal, 15,054 osteopenia, 10,036 Osteoporosis)
Number of testing data	60 (22 normal and Osteopenia, 38 osteoporosis)	54 (11 normal, 22 Osteopenia, 21 osteoporosis) Augmented: 4,069 (939 normal, 1,878 osteopenia 1,252 osteoporosis)	4,069 (939 normal, 1,878 osteopenia 1,252 osteoporosis)
Gold- standard (Classtarget)	Lumbar/Fe-moral BMD DXA (Normal, osteopenia/osteoporosis)	Lumbar/Fe-moral BMD DXA (Normal, osteopenia, osteoporosis)	Lumbar/Fe-moral BMD DXA (Normal, osteopenia, osteoporosis)
Type of Feature extraction	Analysis of morphological (The trabecular area, the periphery, a proportion trabecular length, area, and perimeter (M1-M14)	Analysis morphological (porosity, the size of porosity and orientation of porous)	Texture feature (feature map CNN)
Method of Classification	Classification and Regression Tree Analysis (CART)	C45 (3,6,9-fold validation)	Fully Connected Neural Network (FCNN)
Accuracy	88.33%	86.67%	98.10%
Accuracy from joint of number of testing data [11] and this research	-	-	96.10%
Accuracy from of number testing data [11]	-	-	92.50%

V. CONCLUSION

As shown in Table I, the highest validation accuracy is achieved when the block size is 140x140, the number of convolution layers is 5, and the size of the convolution kernel is 5x5 for the first layer and 3x3 for the other layers. It can then be concluded that the bigger the image block, the higher the validation accuracy. This tells us that bigger images provide additional information that helps discriminate trabecular patterns in normal, osteopenia, and osteoporosis classes. The improvement in accuracy, though, does not change much when the block size is increased from 140x140 to 150x150. This indicates that 140x140 has provided most of the information required by CNNs to distinguish osteoporosis. The training and validation accuracy achieved by the best model is 99.50% and 98.10%, respectively, while the loss of training and validation is 1.30% and 5.40%. Then the testing accuracy is 92.50%.

VI. FUTURE WORK

For this reason, research on osteoporosis examination using dental periapical radiograph images can continue to be carried out, considering that research for osteoporosis examination using dental periapical radiograph images is still rarely used compared to dental panoramic images. This study can be developed by adding a process to increase the

resolution of dental periapical radiographs that tend to be a low resolution at the pre-processing stage and applying the automatic ROI selection method [29].

Further, it can also increase the number of data collections for normal and osteoporosis classes and can use variations in the image of the trabecular bone area of the left and right posterior mandibles.

ACKNOWLEDGMENT

The authors would like to thank the Universitas Gadjah Mada and the Universitas Bumigora Mataram for funding this research. The author also would thank the Radiology Department, Prof. Soedomo Dental Hospital of Universitas Gadjah Mada and the Radiology Department of Dr Sardjito Hospital Indonesia. for supporting this research in acquisition image.

REFERENCES

- [1] M. L. Brandi, "Microarchitecture, the key to bone quality," *Rheumatology*, vol. 48, no. suppl 4, pp. iv3-iv8, Oct. 2009.
- [2] J. E. Adams, "Dual-energy X-ray absorptiometry," *Med Radiol Diagnostic imaging Radiat Oncol.*, vol. 2, pp. 105-124, 2008.
- [3] P. Watanabe et al., "Morphodigital study of the mandibular trabecular bone in panoramic radiographs," *Int. J. Morphol.*, vol. 25(4), no. 4, pp. 875-880, 2007.

- [4] G. D. Elia, G. Caracchini, and G. D. Elia, "P22461252.pdf," Bone Fragility imaging Tech. Clin. Cases Miner. Bone Metab., vol. 6, no. 3, pp. 234–246, 2009.
- [5] K. Horner and H. Devlin, "The relationship between mandibular bone mineral density and panoramic radiographic measurements.," J. Dent., vol. 26, no. 4, pp. 337–343, 1998.
- [6] A. F. Leite, P. T. de S. Figueiredo, C. M. Guia, N. S. Melo, and A. P. de Paula, "Correlations between seven panoramic radiomorphometric indices and bone mineral density in postmenopausal women," Oral Surgery, Oral Med. Oral Pathol. Oral Radiol. Endodontology, vol. 109, no. 3, pp. 449–456, 2010.
- [7] E. Whaites, Essential of dental radiography and radiology, 4th ed. London: Churchill Livingstone Elsevier, 2007.
- [8] R. Licks, V. Licks, F. Ourique, H. R. Bittencourt, and V. Fontanella, "Development of a prediction tool for low bone mass based on clinical data and periapical radiography," Dentomaxillofacial Radiol., vol. 39, no. 4, pp. 224–230, 2010.
- [9] W. G. M. Geraets et al., "Selecting regions of interest on intraoral radiographs for the prediction of bone mineral density," Dentomaxillofacial Radiol., vol. 37, no. 7, pp. 375–379, Oct. 2008.
- [10] W. G. M. Geraets, C. Lindh, and H. Verheij, "Sparseness of the trabecular pattern on dental radiographs: Visual assessment compared with semi-automated measurements," Br. J. Radiol., vol. 85, no. 1016, pp. 455–460, 2012.
- [11] E. I. Sela, S. Hartati, A. Harjoko, R. Wardoyo, and M. Mudjosemedi, "Feature Selection of the Combination of Porous Trabecular with Anthropometric Features for Osteoporosis Screening," Int. J. Electr. Comput. Eng., vol. 5, no. 1, pp. 78–83, 2015.
- [12] E. I. Sela and R. Widyaningrum, "Osteoporosis Detection using Important Shape-Based Features of the Porous Trabecular Bone on the Dental X-Ray Images," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 9, pp. 247–250, 2015.
- [13] E. I. Sela and R. Pulungan, "Osteoporosis identification based on the validated trabecular area on digital dental radiographic images," Procedia Comput. Sci., vol. 157, pp. 282–289, 2019.
- [14] S. F. Diba, R. S. Gracea, D. Rurie Ratna Shantiningsih, and Khasnur Hidjah, "Analysis of mandible trabecular structure using digital periapical radiographs to assess low bone quality in postmenopausal women," Saudi Dent. J., 2021.
- [15] M. M. S. Enny Itje Sela, Sri Hartati, Agus Harjoko Retantyo Wardoyo, "Segmentation on the Dental Periapical X-Ray Images for Osteoporosis Screening," IJACSA, vol. 4, No. 7, p. 158, 2013.
- [16] P. L. Lin, H. C. Hsu, P. Y. Huang, P. W. Huang, and P. Chen, "Alveolar bone-loss area localization in periapical radiographs by texture analysis based on fBm model and GLC matrix," 2014 IEEE Int. Symp. Bioelectron. Bioinformatics, IEEE ISBB 2014, 2014.
- [17] P. Huang, P. Huang, P. Lin, and H. Hsu, "Alveolar bone-loss area detection in periodontitis radiographs using hybrid of intensity and texture analyzed based on fBm model," pp. 13–16, 2014.
- [18] N. Tajbakhsh et al., "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," IEEE Trans. Med. Imaging, vol. 35, no. 5, pp. 1299–1312, May 2016.
- [19] Bakhshwain and A. Sagheer, "Online Tuning of Hyperparameters in Deep LSTM for Time Series Applications," Int. J. Intell. Eng. Syst., vol. 14, no. 1, pp. 212–220, 2020.
- [20] [M. E. Wibowo, "Rancangan Pelatihan Paralel Jaringan Saraf Deep Learning Berbasis Map-Reduce," in Seminar Nasional Teknologi Informasi, Komunikasi dan Industri (SNTIKI) 9 ISSN (Printed) : 2579-7271 Fakultas Sains dan Teknologi, UIN Sultan Syarif Kasim.
- [21] D. D. Himabindu, "A Survey on Computer Vision Architectures for Large Scale Image Classification using Deep Learning," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, pp. 105–120, 2021.
- [22] K. He, X. Zhang, S. Ren, and Jian Sun, "Deep residual learning for image recognition," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 770–778, 2016.
- [23] J. Z. Cheng et al., "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans," Sci. Rep., vol. 6, no. October 2015, pp. 1–13, 2016.
- [24] K. L. Hua, C. H. Hsu, S. C. Hidayati, W. H. Cheng, and Y. J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," Onco. Targets. Ther., vol. 8, pp. 2015–2022, 2015.
- [25] L. Mangeri, "Chest Diseases Prediction from X-ray Images using CNN Models: A Study," Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, pp. 236–243, 2021.
- [26] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.
- [27] M. A. Nielsen, "Neural_Networks_and_Deep_Learning_Pdf.Pdf," Determ. Press, 2015.
- [28] D. Stathakis, "How many hidden layers and nodes?," Int. J. Remote Sens., vol. 30, no. 8, pp. 2133–2147, 2009.
- [29] E. I. Sela, R. Pulungan, R. Widyaningrum, and R. R. Shantiningsih, "Method for automated selection of the trabecular area in digital periapical radiographic images using morphological operations," Healthc. Inform. Res., vol. 25, no. 3, pp. 193–200, 2019.

Prediction of Diabetic Obese Patients using Fuzzy KNN Classifier based on Expectation Maximization, PCA and SMOTE Algorithms

Ibrahim Eldesouky Fattoh¹

Computer Science Dept., Faculty of Computers and
Artificial Intelligence, Beni-Suef University
Beni-suef, Egypt

Soha Safwat²

Software Engineering and Information Technology Dept
The Egyptian Chinese University
Cairo, Egypt

Abstract—Diabetes is a long-term disease. Inappropriate blood sugar level control in diabetic patients can lead to serious issues like kidney and heart diseases. Obesity is widely regarded as a major risk factor for type 2 diabetes. In this research, a model proposed to predict diabetic obese patients based on Expectation Maximization, PCA, and SMOTE Algorithms in the preprocessing and feature extraction phases, and using Fuzzy KNN classifier in the prediction phase. The model applied on real dataset and the accuracy of prediction results reflects the positive effect of the preprocessing techniques. The accuracy of the proposed model is 95.97% and outperforms other model applied on the same dataset.

Keywords—KNN classifier; SMOTE; PCA; diabetic obese patients

I. INTRODUCTION

Using Data Mining (DM) and Machine Learning (ML) techniques in data mining research are a common way for making use of large amounts of available knowledge-based data. Machine Learning is extremely essential in the realm of medical diagnostics. Data mining is a great goal in science and medical research, which necessarily generates massive amounts of data owing to the special societal effect of the serious disease. As a result, Machine learning and data mining approaches are unquestionably of great importance for aspects of clinical administration, diagnosis, and treatment. As part of this work, challenges were undertaken to examine the recent literature on ML and DM methodologies in many diseases especially in the diseases of the chronic diabetes. Diagnosis in the healthcare sector is an ideal subject for ML algorithms [1]. Many of these may be identified using pattern recognition on large amounts of data. An algorithm should be trained on a small number of tests to be useful in the field, medical diagnostics must be able to tolerate noisy and empty datasets. Many researches on machine learning in the sector of healthcare have been undertaken. Healthcare ML has emerged as a top goal for many academics. Different data mining approaches and procedures in hidden pattern recognition can be used to gain insights. The medical science primary roles are to prevent or help treat diseases. One of the chronic illnesses marked by hyperglycemia is Diabetes mellitus. It can lead to a slew of difficulties [2]. As a result of higher mortality rates in recent years, According to WHO (World Health Organization) forecasts by 2040, the world's population of diabetes is

anticipated to reach 642 million [3], suggesting that one out of every ten people would suffer from diabetes in the future. There are three types of Diabetes [4], namely; Gestational Diabetes, Type-1 Diabetes Mellitus, Type-2 Diabetes Mellitus. Type-2 diabetes mellitus patients are frequently classified as having a fatty liver disease in which it could be either nonalcoholic or alcoholic fatty liver disease (NAFLD|AFLD) [5]. Type-2 diabetes mellitus has been postulated as a primary cause of NAFLD development, or nonalcoholic steatohepatitis, which likely reflects in Type-2 diabetes mellitus with rapid advancement of weight gain and resistance of insulin. Obesity and diabetes, both multifactorial, difficult illnesses, have become major public health issues across the world [6]. Many conditions, on the other hand, may be prevented. Obesity is a significant growing health concern; some refer to it as the New World Syndrome [7]. The occurrence of obesity and fatty liver in persons with diabetes of Type-2 has long been documented as they are strongly associated with each other [8]. It is often viewed as an accidental finding with small to no therapeutic value. Sedentary lifestyles or poor dietary habits result in weight gain. It may also increase the chances of facing a metabolic syndrome over time. Avoiding the significant consequences that result in massive issues in health, since early detection is the beginning point for a good life without the disease reflects the significance of using the recommended method for predicting patients suffer from diabetes and affected by obesity and NAFLD. Diabetes mellitus and its consequences, in particular, must be prevented and managed in poor and middle-income countries. The following is how this paper is arranged; Section II outlines the related work. Section III, details the suggested model as well as the dataset used. The Section IV offers the obtained results, followed by the conclusion and the future work in Section V.

II. RELATED WORK

In [9], Kumar purposed various data mining approaches in medical sector to highlight data mining applications based on the nature of the information; In order to predict Parkinson's illness, Support Vector Machines and Artificial Neural Networks were used and resulted in 95 percent accuracy. In addition, it improved detection rate by employing an ANN to diagnose cancer of breast to 98.8 percent, and employed Artificial Neural Networks. Basma Boukenze et al. in [10] assessed the DM techniques performance in medical health

sector using multiple learning techniques. The result simulation indicated that the decision tree (DT) performed better than other learning techniques in forecasting kidney failure chronic disease. Furthermore, M. Abdullah and S. Al-Asmari in [11], clarified the same DM approaches to designate the type of anemia patients suffer from anemia. DT executed with an accuracy result of 93.75 percent. While only support vector machine algorithm was used in categorizing diabetes disease, while in [12] Kumari and Chitra used the Matlab tool version 2010a in order to identify the diabetic patients by 78 percent accuracy. Developing DT and DM classification approaches assists medical practitioners in gaining better medical judgments to detect diseases timely [13]. El-Halees and Shurrab in [14] developed a model that can discriminate between individuals with blood tumors and normal blood illness utilizing multiple association rules and ANN, results with 79.45 percent accuracy. In addition, in order to predict diabetes in many circumstances various researches have been conducted in which the authors of [15] used a regression-based approach of DM to introduce diabetes therapy predictive analysis. The Oracle Data Miner was used as a mining software to forecast diabetic treatment methods. For the experimental investigation, the support vector machine technique was applied. They conclude that pharmacological therapy for patients under the age of 18 can be postponed to minimize negative effects. The authors used four classifiers in [16] to categorize the diabetes mellitus risk. First, four categorization models were investigated: DT, Logistic Regression, ANN and Naive Bayes. Then, to improve the resilience of such models, Bagging and Boosting strategies were examined. According to the findings, the Random Forest (RF) algorithm performs the best in illness risk categorization. They suggested an early diabetes prediction model in [17], and they discovered a high correlation between diabetes, glucose level and body mass index (BMI), that was retrieved using the Apriori technique. Diabetes was predicted using RF, ANN, and K-means approaches. The ANN approach achieved the highest accuracy of 75.7 percent. For the prediction of diabetes, the authors of [18] employed KNN and the Naïve Bayes approach. Their method was implemented as a program of expert software, in which users submit input in the form of patient data and the determination of whether or not the patient is diabetic. The authors of [19] propose an attribute selection technique of firefly and cuckoo search-based for the PIMA Indian diabetes database from University of California Irvine (UCI), with the goal of greater accuracy and lesser training overhead. They also said that the proposed structure promises to be more accurate than the usual technique. The authors of [20] applied a ML model to forecast the occurrence of Type-2 Diabetes mellitus, using information from the present year (Y). From 2013 to 2018, electronic health records were collected at a private medical facility for this investigation. Key characteristics were initially picked for the prediction model using chi-squared tests, ANOVA tests and recursive variable reduction approaches. Based on these variables, they used random forest, logistic regression, XGBoost, SVM and ensemble ML methods in order to foresee the result as diabetic, non-diabetic or pre-diabetic. The model performed pretty well in anticipating the occurrence of Type-2 diabetes in the Korean population. The authors of [21] applied two machine-learning

techniques for two-phase classification; SVM and ANN to predict diabetes mellitus. They used a real dataset from Al-Kasr Al-Aini Hospital in Egypt. In the first phase, they used SVM to predict patients with fatty liver disease with accuracy of 95%. Then in the second phase they used ANN for prediction of diabetic patients based on phase 1 and another 8 different attributes.

III. PROPOSED SOLUTION AND DATASET

As the dataset of this problem was collected manually as will be described in next section, it had many issues like missing values, and the data was unbalanced, so we applied a preprocessing phase for the dataset. The algorithms used in the proposed model are described in this section.

A. Expectation Maximization Algorithm for Estimating the Missing Values

Dempster et al. 1977 in [22], demonstrated that the Expectation Maximization (EM) algorithm can be applied when X_{MIS} (the missing data joint distribution) and X_{OBS} (the observed data) is candid. For all $(\theta \in Rd)$, let the density function probability of $(;)$ be $X=(X_{OBS}, X_{MIS})$. The estimation of θ get the most out of the observed data log eventuality in which the expectation maximization algorithm aims to find.

$$(\theta; X_{OBS}) = \log(X_{OBS}; \theta) = \log \int f(X_{OBS}, X_{MIS}; \theta) dX_{MIS} \quad (1)$$

In general, because this number cannot be estimated explicitly, the EM method calculates the MLE by iteratively maximizing the anticipated log-likelihood of complete-data in (2)

$$l(X; \theta) = \log f(X_{OBS}, X_{MIS}; \theta) \quad (2)$$

Begin with a value of θ^0 and let θ^t be the estimate of at the t^{th} iteration, then below is two steps of the next EM iteration:

E step: Calculate the expectation of log-likelihood of complete-data in relation to the missing covariate conditional distribution parameterized by (θ^t) :

$$Q(\theta, \theta^{(t)} = E[l(X; \theta) | X_{OBS}; \theta^{(t)}] = \int l(X; \theta) f(X_{MIS} | X_{OBS}; \theta^{(t)}) dX_{MIS} \quad (3)$$

M step: Define $((t+1))$ by maximizing the Q function:

$$\theta^{(t+1)} \in \text{argmax}_{\theta} (Q(\theta, \theta^{(t)})) \quad (4)$$

B. Feature Reduction using Principal Component Analysis Algorithm

Principle Component Analysis (PCA) is an extracting features statistical approach that employs to turn a set of possibly associated annotations to a set of variables uncorrelated transformed linearly known as principle components. PCA may be used to reduce the feature dimensions [23]. Because the eigenvectors number exceeds the columns number, the dimension of the projected output data is smaller than the dimension of the input data. The method of PCA utilized in the feature reduction step is as follows.

Algorithm 1 " PCA"

Assume: samples of N (x_1, \dots, x_N) each $x_n \in R^D$
 Aim: D dimensions data project to K dimensions ($K < D$)
 Then captures the projected data maximum possible variance

Consider

(u_1, \dots, u_D) be the principle components, assumed to be
 orthogonal such that: $u_i^T u_j = 0$ if $i \neq j$ and should be
 orthonormal such that $u_i^T u_i = 1$

Each peincipal component is a vector of size $D \times 1$

We will select the frirst K principal components

1: Compute the mean of the data

$$\bar{X} = \frac{1}{N} \sum_{n=1}^N X_n \quad (5)$$

2: Compute the Covariance matrix

$$s = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T \quad (6)$$

3: Find the eign values 4

4: Take the top k eign vectors according
 (corresponding to the top k eign values)

5: Call these vectors as u_1, u_2, \dots, u_k

(such that $\lambda_1 \geq \lambda_2 \dots \dots \geq \lambda_{k-1} \geq \lambda_k$

$$6: Z_n = U^T \times X_n \quad (7)$$

C. Handling the Un-Balanced Data using SMOTE Algorithm

Chawla presented the Synthetic Minority Oversampling Technique (SMOTE) in 2002 [24]. In contrast to random oversampling, in the SMOTE method the minority class is oversampled by producing samples of synthetic rather than oversampling with replacement. The SMOTE method generates fake instances based on similarities between existing minority cases in feature space rather than data space [24, 25]. These synthetic instances are constructed by connecting a portion or all of the minority class's K-Nearest Neighbor (KNN). Neighbors from the KNN are picked at random depending on the quantity of oversampling necessary. Algorithm 2 represents the used SMOTE algorithm for handling the un-balanced dataset.

D. Fuzzy KNN Classifier

Keller et al introduced the fuzzy KNN classifier[26], which assign to each sample a class memberships, as a function from of its KNN training samples of each sample's distance. Because it is easy and provides information on the certainty of the classification result, the fuzzy KNN classifier is a popular choice for applications. According to Keller et al, the major benefit of utilizing the FKNN model may not be the reduction in error rate. More crucially, the model provides a level of assurance that may be combined with the "refuse-to-decide" option. Objects with overlapping classes can thus be discovered and treated independently as in Algorithm 3.

Algorithm 2 " SMOTE"

The input:

X: original set of training sample

N: percentage of oversampling

K: nearest neighbors value

The output: the oversampled training set

n ← # observations

m ← # attributes

nmin ← # min observations

if N < 100 then

Stop: warning "N should be greater than 100"

end if

N ← int(N/100)

S_{(n*N)*m} ← empty array for synesthetic samples

for i ← 1 → n_{min}

Do

Discover the KNN for each i then save the indexes in the m newindex ← 1

while N ≠ 0 do

K_c ← number between (1& K) randomly

for j ← 1 → m

do

$$\text{dif } f \leftarrow X[\text{nn}[K_c][j]] - X[i][j] \quad (8)$$

gap ← uniform(0, 1)

$$\text{synthetic}[\text{newindex}][j] \leftarrow X[i][j] + \text{gap} \times \text{dif } f \quad (9)$$

end for

newindex+ = 1

N- = 1

end while

end for

Return (X & synthetic)

Algorithm 3 " Fuzzy K-Nearest Neighbor

From sample x, get the KNN.

Soft labels, input x, a membership vector

$$\mu(x) = [\mu_{c_1}(x), \dots, \mu_{c_i}(x), \dots, \mu_{c_l}(x)]$$

$$\mu_{ih}(x) = \mu_i(x_i) = \begin{cases} 0.51 + \left(\frac{n_i}{k}\right) * 0.49, & \text{if } c(x_j) = i \\ \left(\frac{n_i}{k}\right) * 0.49, & \text{if } c(x_j) \neq i \end{cases} \quad (10)$$

- Membership function calculation

$$\mu_i(x) = \frac{\sum_{j=1}^k \mu_{ij} (1 / (\|x - x_j\|^{2/(m-1)}))}{\sum_{j=1}^k (1 / (\|x - x_j\|^{2/(m-1)}))} \quad (11)$$

- The target class $\max \mu_i(x)$

E. Dataset Description

The dataset used in this study was obtained from Cairo University, Faculty of Medicine, Al-Kasr Al-Aini Hospital [21]. The dataset contains 30 variables; Gender, Age, Alcohol consumption, Smoking, Schistosomiasis, steroids, History of hypertension, Oral contraceptive pill, Waist circumference, Body Mass Index, Hemoglobin test (HGB), Liver disease, Primed lymphocyte test, Basic Insulation Level, Aspartate Aminotransferase (AST), Alanine Aminotransferase (ALT), White blood cells (WBCs), Albumin level in blood (ALB), Protein C test, Alkaline phosphatase (ALP), Gamma-Glutamyl Transferase (GGT), Total cholesterol, Triglycerides test (TGs), High-density lipoprotein (HDL), Low-density lipoprotein (LDL), International Normalized Ratio (INR), Spleen size, Fasting blood sugar, History of diabetes, and Hemoglobin A1c (HBA1C). This was preprocessed as will be shown in the proposed model section through different phases. The algorithms used in the data-preprocessing phase are expectation maximization algorithm, which estimate missing values. PCA algorithm is used in feature reduction phase, while SMOTE algorithm used to generate new sample in the minority class to overcome the unbalanced data issue that affects the measures.

F. Proposed Model

Fig. 1 shows the basic steps used in the proposed model for the prediction of diabetic obese patients. At the first, we read the dataset and apply a data preprocessing phase on it. The first step in the data preprocessing is estimating the missing values by the EM algorithm. The next step is applying the PCA algorithm to reduce the features in the dataset. The basic steps of the PCA are calculating the covariance matrix, then calculating the Eigen values, then sorting the attributes in descending order, then normalizing the values, and calculating the weight value for each attribute. The third step is solving the unbalanced data using SMOTE algorithm described in last section above. The SMOTE algorithm used for generating new sample in the minority class. The last step in the proposed model is classifying the new samples using fuzzy KNN classifier.

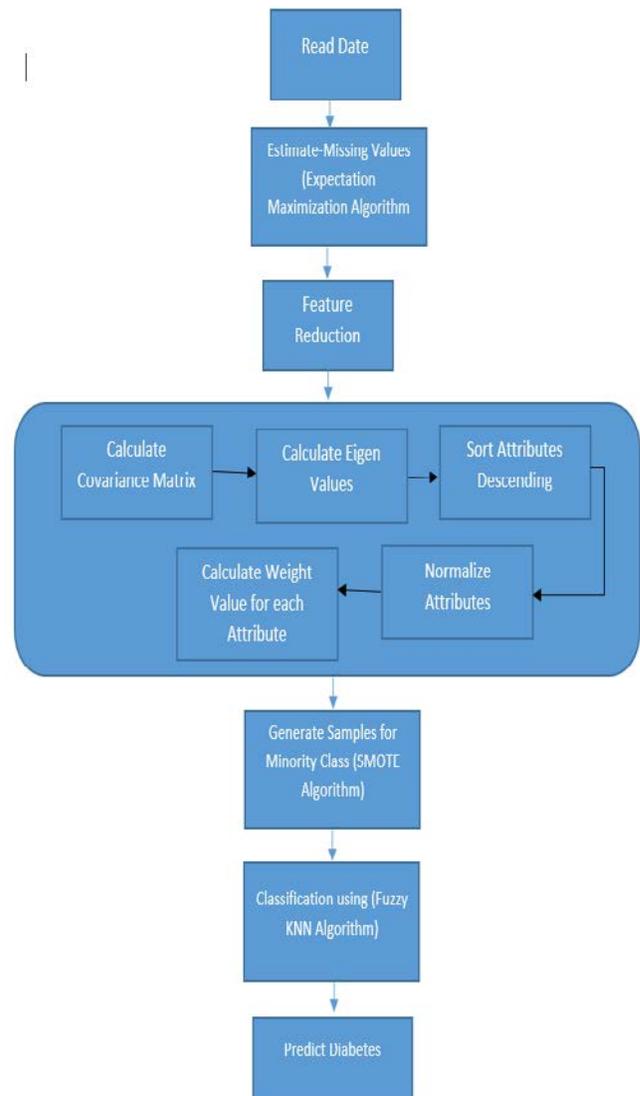


Fig. 1. Proposed Model Steps for Prediction of Diabetic Obese Patients.

IV. EVALUATION METHODS AND RESULTS

A. Evaluation Method

Evaluation of the performance of algorithms using the precision and recall criteria is very valuable. When making a choice, precision is the proportion of the time that the model properly predicts a good outcome. Precision is defined as the accurately identified or predicted positive examples divided by all the positive examples given. The proportion of properly recognized positives out of all existing positives is referred to as recall; it is calculated by dividing by the truly categorized positive cases by all the number of genuine examples in the set of positive testing. An optimal model must have both high recall and great accuracy. The F-measure is the consistent measure of accuracy and recall. The F-measure runs from zero to one, in which one indicating a classifier that properly captures accuracy and recall.

$$Precision = \frac{TP}{TP+FP} \tag{12}$$

$$Sensitivity = \frac{TP}{TP+FN}, \tag{13}$$

$$F - measure = \frac{2(Precision)(Sensitivity)}{(Precision)+Sensitivity} \tag{14}$$

In which, the true positive (TP) represent the positive cases that predicted positive, the false negative (FN) represents the cases that were positive. However, it predicted negative and the false positive (FP) are the negative cases that were positively predicted.

B. Results

In this part, we report the findings obtained when the fuzzy KNN classifier used with the proposed model on the dataset described, and applying the fuzzy KNN classifier on the raw data of the dataset. Table I shows the proposed model output applied on the dataset after preprocessing compared to the same classifier but without data preprocessing.

TABLE I. FUZZY KNN (PROPOSED) WITH DATA PREPROCESSING COMPARED TO FUZZY KNN WITHOUT DATA PREPROCESSING

	FKNN Proposed Model Fuzzy Parameter = 0.5 & K= 5	F-KNN Raw data Fuzzy Parameter = 0.5 & K= 5
Sensitivity	1.0000	0.7156
Precision	0.9197	0.5735
Accuracy	0.9579	0.6577
F1 Score	0.9582	0.6367

Table I and Fig. 2 shows that the data preprocessing steps, estimating the missing values, feature reduction and solving the problem of unbalanced data enhanced the all measurement values resulted from the classifier.

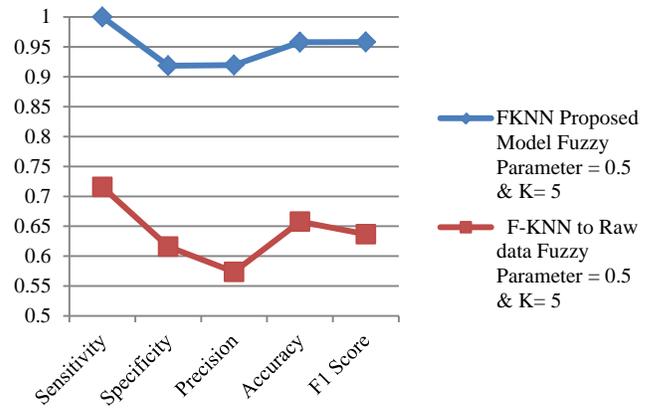


Fig. 2. Proposed Model with Data Preprocessing Vs Fuzzy KNN without Data Preprocessing.

The obtained results were compared to the results obtained in [21], as they used the same dataset. They proposed a two-phase classifier for predicting the potential diabetic obese patient as mentioned in related work section. Table II shows the basic differences in the proposed model and the model in [21].

TABLE II. THE DIFFERENCE IN METHODOLOGIES USED IN THE PROPOSED MODEL WITH THE MODEL IN [21]

	Proposed model	Model in [21]
Estimating missing values	Expectation Maximization Algorithm	weighted sum of linear interpolations from the closest accessible points.
Feature reduction	PCA algorithm	Correlation Feature Selection
Handling unbalanced data	SMOTE algorithm	Using K-fold cross validation
Classification	Fuzzy KNN classifier	SVM for phase1 and ANN for phase 2

Table III shows the comparison between results of the proposed model and results in [21].

TABLE III. ACCURACY COMPARISON BETWEEN PROPOSED MODEL AND MODEL IN [21]

	Proposed model	Model in [21]
Accuracy	95.97%	86.56%

From Tables II and III, we can observe that the algorithms and techniques used in the proposed model to prepare the data before training and testing were affected positively the data especially the steps of estimating missing values and handling unbalanced data, also the proposed classifier introduces a promising classification accuracy compared to the results introduced in [21].

V. CONCLUSION AND FUTURE WORK

In this research a model for prediction of Diabetic Obese Patients was proposed, the model was based on Expectation Maximization, PCA, and SMOTE Algorithms in data preparation and preprocessing phase, and the fuzzy KNN classifier was used in prediction phase. The dataset used in this research was obtained from Cairo University, Faculty of Medicine, Al-Kasr Al-Aini Hospital. The algorithms used in the preprocessing enriched the clearness and effectiveness of the dataset which reflected in the prediction phase as shown in the results. The prediction accuracy reached to 95.97% in the proposed model and this result outperforms a corresponding model applied on the same dataset mentioned in the related work. We can suggest some improvements in the preprocessing phase afterwards like adopting another feature selection algorithm and other algorithms for handling imbalanced data, and estimating the missing values. In addition, an ensemble model can be provided on more than one classifier in order to enhance the precision value.

REFERENCES

- [1] Nilashi M, bin Ibrahim O, Ahmadi H, Shahmoradi L. An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering* 2017;106:212-23.
- [2] Cichosz SL. Predictive models in diabetes: Early prediction and detecting of type 2 diabetes and related complications: Aalborg Universitetsforlag; 2016.
- [3] Zou Q, Qu K, Ju Y, Tang H, Luo Y, Yin D. Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics* 2018;9:515.
- [4] Devi MR, Shyla JM. Analysis of various data mining techniques to predict diabetes mellitus. *International Journal of Applied Engineering Research* 2016;11:727-30.
- [5] Mills EP, Brown KPD, Smith JD, Vang PW, Trotta K. Treating nonalcoholic fatty liver disease in patients with type 2 diabetes mellitus: a review of efficacy and safety. *Therapeutic advances in endocrinology and metabolism* 2018;9:15-28.
- [6] Bhupathiraju SN, Hu FB. Epidemiology of obesity and diabetes and their cardiovascular complications. *Circulation research* 2016;118:1723-35.
- [7] Ng M, Fleming T, Robinson M, Thomson B, Graetz N, Margono C, et al. Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The lancet* 2014;384:766-81.
- [8] Ballestri S, Zona S, Targher G, Romagnoli D, Baldelli E, Nascimbeni F, et al. Nonalcoholic fatty liver disease is associated with an almost twofold increased risk of incident type 2 diabetes and metabolic syndrome. Evidence from a systematic review and meta-analysis. *Journal of gastroenterology and hepatology* 2016;31:936-44.
- [9] Kumar RN, Kumar MA. Medical Data Mining Techniques for Health Care Systems. *International Journal of Engineering Science* 2016;3498.
- [10] Boukenze B, Mousannif H, Haqiq A. Performance of data mining techniques to predict in healthcare case study: chronic kidney failure disease. *Int Journal of Database Management systems* 2016;8:1-9.
- [11] Abdullah M, Al-Asmari S. Anemia types prediction based on data mining classification algorithms. *Communication, Management and Information Technology—Sampaio de Alencar* (Ed) 2017.
- [12] Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications* 2013;3:1797-801.
- [13] Daghistani T, Alshammari R. Diagnosis of diabetes by applying data mining classification techniques. *International Journal of Advanced Computer Science and Applications* (IJACSA) 2016;7:329-32.
- [14] El-Halees, A. M., & Shurrah, A. H. (2017). Blood tumor prediction using data mining techniques. *Health Informatics—An International Journal*, 6.
- [15] Aljumah, A. A., Ahamad, M. G., & Siddiqui, M. K. (2013). Application of data mining: Diabetes health care in young and old patients. *Journal of King Saud University-Computer and Information Sciences*, 25(2), 127-136.
- [16] Nai-arun, N., & Moungrmai, R. (2015). Comparison of classifiers for the risk of diabetes prediction. *Procedia Computer Science*, 69, 132-142.
- [17] Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., ... & Abbas, Z. (2019). A model for early prediction of diabetes. *Informatics in Medicine Unlocked*, 16, 100204.
- [18] Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining. *Innovations in information, embedded and communication systems (ICIIECS)*, 2017 international conference on. 2017. p. 1–5.
- [19] Haritha, R., Babu, D. S., & Sammulal, P. (2018). A hybrid approach for prediction of type-1 and type-2 diabetes using firefly and cuckoo search algorithms. *International Journal of Applied Engineering Research*, 13(2), 896-907.
- [20] Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int. J. Environ. Res. Public Health* 2021, 18, 3317. <https://doi.org/10.3390/ijerph18063317>.
- [21] Ali, R. E., El-Kadi, H., Labib, S. S., & Saad, Y. I. (2019). Prediction of potential-diabetic obese-patients using machine learning techniques. (IJACSA) *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 8, 2019.
- [22] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.
- [23] Lakhina, S., Joseph, S., & Verma, B. (2010). Feature reduction using principal component analysis for effective anomaly-based intrusion detection on NSL-KDD. *International Journal of Engineering Science and Technology* Vol. 2(6), 1790-1799.
- [24] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16(1):321–357, 2002.
- [25] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *Knowledge and Data Engineering, IEEE Transactions on*, 21(9):1263–1284, 2009.
- [26] Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4), 580-585.

Evaluation Optimal Prediction Performance of MLMs on High-volatile Financial Market Data

Yao HongXing¹, Hafiz Muhammad Naveed^{2*}

Muhammad Usman Answer³, Bilal Ahmed Memon⁴, Muhammad Akhtar⁵

School of Finance and Economics, Jiangsu University, Zhenjiang, China^{1,2,5}

School of Computer Engineering, National University of Computer and Engineering Sciences, Lahore, Pakistan³

School of Business Administration, Iqra University, Karachi, Pakistan⁴

Abstract—The present study evaluates the prediction performance of the multi-machine learning models (MLMs) on high-volatile financial markets data sets since 2007 to 2020. The linear and nonlinear empirical data sets are comprised on stock price returns of Karachi stock exchange (KSE) 100-Index of Pakistan and currencies exchange rates of Pakistani Rupees (PKR) against five major currencies (USD, Euro, GBP, CHF & JPY). In the present study, the support vector regression (SVR), random forest (RF), and machine learning-linear regression model (ML-LRM) are under-evaluated for comparative prediction performance. Moreover, the findings demonstrated that the SVR comparatively gives optimal prediction performance on group1. Similarly, the RF relatively gives the best prediction performance on group2. The findings of study concludes that the algorithm of RF is most appropriate for nonlinear approximation/evaluation and the algorithm of SVR is most useful for high-frequency time-series data estimation. The present study is contributed by exploring comparative enthusiastic/optimistic machine learning model on multi-nature data sets. This empirical study would be helpful for finance and machine-learning pupils, data analysts and researchers, especially for those who are deploying machine-learning approaches for financial analysis.

Keywords—Support vector regression; random forest; machine learning-linear regression model; optimal prediction performance; currencies exchange rates; stock price returns

I. INTRODUCTION

Since several decades, a lot of researchers and data analysts have been applying traditional econometric models (TEM) for hypothetical testing and financial-nonfinancial market evaluation. But there have some constraints with TEM as nonlinear approximation, future prediction developments of the markets and data prediction accuracy. On the other hand, nowadays, the supervised and unsupervised machine learning approaches have quite famous to precisely evaluate the different nature of big data. So, the analysis effectiveness is directly associated with the prediction accuracy the model.

The machine learning approaches are successfully employed into different domains with respect to their applications. But it is usually used for speech recognition, image processing, wind-speed frequency scaling, network filtering and financial markets prediction [1-4]. Furthermore, the evaluation of stock market and forex market returns with machine learning approaches are very hot topic nowadays.

Moreover, the financial-nonfinancial future markets predictions are very helpful for those who are willing for financing in the market and for hedgers to hedge their financial assets [5, 6]. The machine learning algorithms (MLAs) have capability to analyze the linear-nonlinear data [7]. For example, the forex market is very dynamical market due to globalization. In prior era, the researchers and data analysts were used TEM to examine the forex market returns. But since last decade, the researchers and data experts have been rapidly diversifying from TEM to machine learning approaches for financial market prediction [8-10]. In fact, the MLMs are most powerful and very effective approaches to predict the high-volatile and nonlinear data financial data [11, 12]. So, the key purpose of present study is to evaluate the prediction performance of underlying machine learning models on different nature financial data sets, such as the group1 comprised on stock price returns of KSE 100-index of Pakistan which is illustrated in Fig. 1.

Fig. 1 indicates that the stock price returns of KSE 100-index is very dynamic financial market. This diagram is comprised on original data, rolling mean and average which collectively demonstrates the good returns (positive returns), bad returns (negative returns) and no returns (zero returns). The group2 contains on the PKR exchange rates against five major currencies (USD, Euro, GBP, CNY & JPY) which is illustrated in Fig. 2.

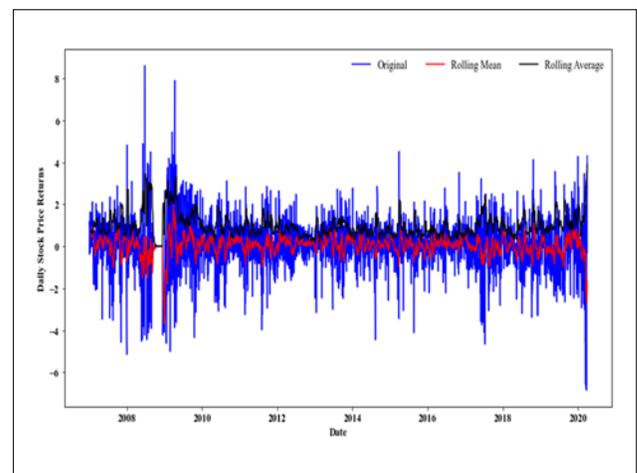


Fig. 1. Daily Stock Price Returns of KSE 100-index.

*Corresponding Author.

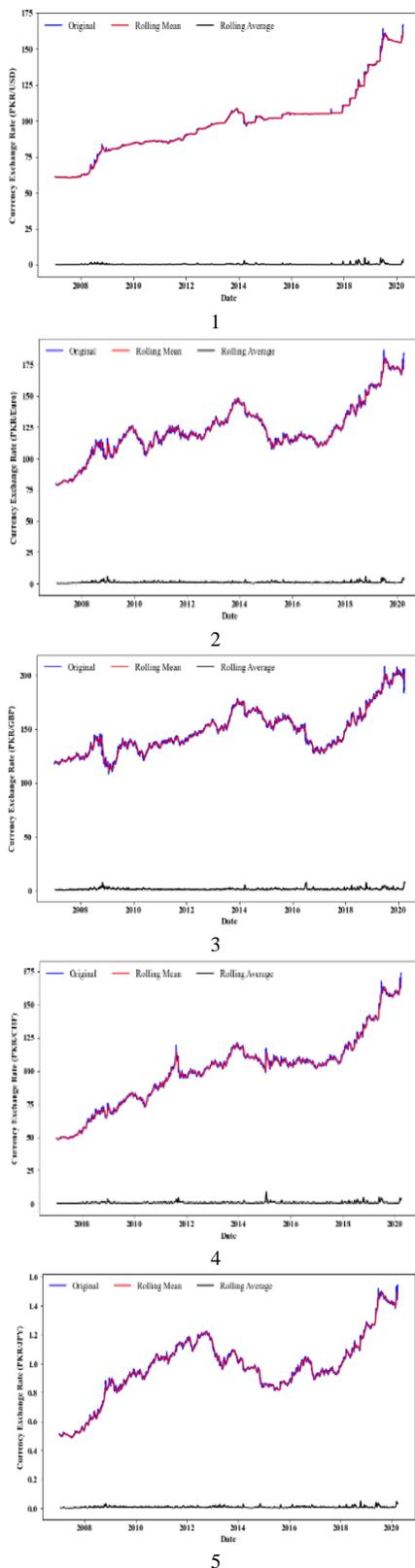


Fig. 2. PKR Exchange Rates against Five Major Currencies.

Fig. 2 demonstrated that the PKR exchange rates against entire currencies have been steadily increasing over the period, which indicating the domestic currency (PKR) has

been depreciating since 2007 to till 2020. Furthermore, these both empirical data sets will be evaluated by SVR, RF and ML-LRM. These empirical machine learning models are commonly used for classification and regression problems. But the present study will be executed the prediction performance of entire machine learning models on nonlinear and high-frequency data sets which quite effective for researchers and data analysts.

The remaining sections of study comprised as: the Section 2 will be contained on most similar literature on the stock markets and the forex markets which executed by MLMs. The Section 3 will be contained on data collection process, data preparation and mathematical interpretation and application of the models. Similarly, the Section 4 will be contained on results and discussion and forecast optimal prediction model. Finally, the Section 5 will be contained on conclusion-based sound evidence, limitation of the study and significant future developments of the study.

II. SIMILAR STUDIES

Since the last decade, the researchers and data analysts have been rapidly diversified from TEM to flexible and experimental machine learning approaches for optimal prediction. In the modern era, the MLAs are typically used to determine the returns of dynamical stock markets. To predict the next day price index of the Taiwan index feature, [13] have deployed hybrid SVR with SOFT (self-organized feature map) approach by filter-based feature selection to improve the prediction accuracy of the model and mitigate the cost of training time. The findings of this study demonstrates that the hybrid SVR improves the training time and prediction accuracy. Another study has employed SVR for prediction stock prices on daily and up to the minute. The findings of this study shows that the SVR has absolute predictive power, whenever applicate the new strategy of model periodically [14]. The hybrid prophet-SVR model is outperformed than relatively other approaches by using time series demand with seasonality in stock index [15]. [16] have applied the machine learning based RF to examine the future predictions of stock prices of key listed firms. Moreover, [17] have been executed the trend of the Thai-stock market by deploying the SVR, multi-layer perception (MLP), and partial least square classifier models (PLSCM). The scholars have concluded that the Thai-stock market is rapidly growing while the investors are absolutely acquaintance about market dynamics. Similarly, some studies have evaluated the factors of stock prices returns via the reinforcement learning approach [18]. A lot of other studies have executed the real impact of prices by using mixed methodologies. Moreover, the Long-Short Term Memory (LSTM) is an artificial recurrent neural network (RNN) architecture which works on feedforward standards. Especially, it employed for outliers' detection problem in given big data set by fixing threshold, hence, this approach is very useful to investigate the dynamical stock market [19-21]. But the present study will execute the stock price returns of KSE 100-index of Pakistan and subsequent evaluate the optimal prediction performance model.

Furthermore, the forex market is quite significant for those who want to be financing into the international market for

economic privileges. Because the currency exchange rates may bring probabilistic effects on economic value of their imperative assets. For currency risk management, a researcher community has been predicted the future currency exchange rates on voluntarily currency selection by using MLMs. For instance, The stochastic volatility model with jumps to SVR in order to account for sudden big changes in exchange rate volatility [22]. So the experimental studies demonstrate that the empirical new model has the ability to improve the forecast accuracy. The author in [23] had predicted the exchange rates of major currencies of developed countries. This study deployed the mixed methodologies as the deep neural networks, MLMs and econometric approaches to predict the currencies exchange rates. Besides, this study concludes that the currencies exchange rates of developing countries are more volatilized than developed countries. So, the traditional models may not properly investigate to the nonlinear approximation. According to econometric assumption, the empirical data should be in linear format for econometric modelling. But the empirical data lost their originality whenever it transformed from nonlinear to linear scale through the natural logarithm or another approach. But the MLMs have ability to evaluate every kind of data. Therefore, the usage of MLAs have been precipitously increasing over the period for optimal prediction to nonlinear approximation. The author in [24] has analyzed the couple of currencies exchange rates (USD/CHF, USD/CAD and USD/JPY) along with the shuffled frog leaping algorithm (SFLA). Moreover, a lot of other scholars executed MLMs prediction performance along with different currencies exchange rates data. Another study executed the currency exchange rate is a certain cause of domestic inflation by using the Dorn-busch approach. Besides, they also investigated that the USA housing prices raised 7% with excluded currency ERs. But in the last decade, the USA housing prices 40% grown-up by variation into local currency exchange rates [25]. After intensely viewed to the comprehensive literature, we summarized that the MLMs are excessively used to predict the stock market and forex market returns.

III. EMPIRICAL DATA AND RESEARCH METHODOLOGY

A. Data Collection and Preparation for Analysis

The empirical data set of the present study is divided into two certain groups. The group1 is contained on the stock price returns of KSE 100-index of Pakistan. In addition, the group1 data is dragged from the official website of the KSE of Pakistan (<https://www.psx.com.pk/>). This study measures the price returns from given stock prices as follow: Lumley [26].

$$R(t_0, t_1) = \left(\frac{P_{t_1} - P_{t_0}}{P_{t_0}} \right) * 100 \quad (1)$$

In Equation 1, the stock price returns are executed as the P_{t_0} is daily stock prices at the time t_0 subtract from the stock price P_{t_1} at time t_1 and divided by P_{t_0} and multiplied by 100, whereas the P_{t_1} denotes to current stock prices and P_{t_0} denotes to the previous stock price of the stock market. As Fig. 1 is showing the stock price returns are high-volatile while this study follows [27-29] for normalizing the empirical data set.

$$y = \ln(r + \sqrt{r^2 + 1}) \quad (2)$$

Moreover, the group2 is contained on PKR exchange rate against five major currencies (USD, EUR, GBP, CHF, JPY). The thirteen-year daily currencies exchange rates data set dragged by the official website of the central bank of Pakistan (<http://www.sbp.org.pk/>). The Fig. 2 discretely exhibits the volatility of currency exchange rates against each given currency.

B. Research Methodology

The certain methodologies play an important role in the empirical studies to execute the hypothetical examination. But before to mathematical interpretation of the models, we must take keenly overview the whole process of performance evaluation of the models. Fig. 3 showed that the primarily both kinds of raw data sets are uploaded into the data repository of jupyter notebook and subsequently refine the data and scaled the data and become in specific format to optimize the machine learning algorithm. Furthermore, the scaled data set would split into train and test, whenever 80% will use for training purpose and rest of data will use for testing purpose. Furthermore, we shall train the model and evaluate the prediction performance of training model with average loss function. This study would execute the prediction performance of the models, if the train predicted error $\xi_t < \xi_e$ expected error otherwise update the training data set.

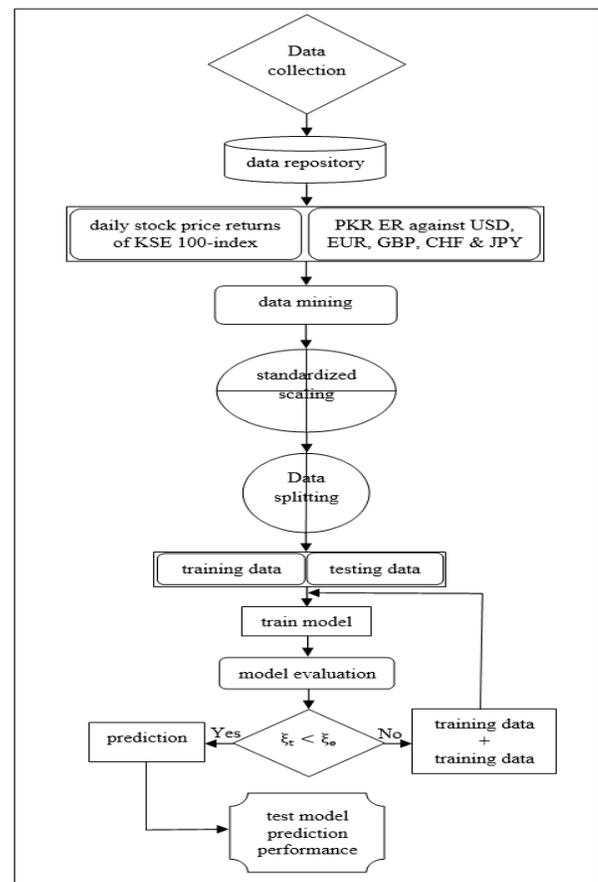


Fig. 3. Flowchart of Models' Prediction Process.

1) *Support vector regression*: The support vector regression model is popular as supervised learning approach which used for sort regression analysis [30, 31]. The support vector machine (SVM) was primarily built for structural risk measurement, but subsequent the researchers and data analysts had customized it for regression purposes, especially for time-series data. The SVM has the ability to measure linear and nonlinear approximation. The authors in [32] have employed the support vector machine (SVM) for the statistical problem. For the execution of the model prediction performance, the present study follows [33, 34].

In Fig. 4, $x_i \in \mathbb{R}_n$ is an input function whereas $i - th$ is $n - th$ number of x as $i = 1, 2, 3, \dots, n$. and $y_i \in \mathbb{R}_n$ is an output function where $i - th$ is $n - th$ number of y as $i = 1, 2, 3, \dots, n$. where $k - th$ is x vector at $n - th$ number of vectors which associated with specific weights $\bar{\alpha}$ at $m - th$ number of $k - th$ vectors and summation b bias. For instance, we have a time-series data set where $z = (x_i y_i), 1 \leq i \leq N$. The core idea dragged by [35]. Let suppose $\bar{\alpha}_m = W^t$, the problem may solve as follows:

$$f(x_i) = W^t \varphi(x_i) + b \tag{3}$$

In equation 3, the w^t represents to specific weight at $n - th$ number of x vector and b denoted to biases. Moreover, x is considered as input vector taken by φ into a higher dimensional space. The model performance can be improved by optimization the weights and bias of the model as below:

$$\min_{(w, b, \varepsilon_1, \varepsilon_2)} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^n (\varepsilon_1^i, \varepsilon_2^i) \tag{4}$$

In equation 4, the C is makes swapping between two models' simplicity and generalization ability. If $Y_i - W^T(\phi(x)) - b \leq \xi + \varepsilon_1^i$ and $W^T(\phi(x)) + b - y_i \leq \xi + \varepsilon_1^i$, the error position is determined by slack variables as ξ & ε_i . We can map high volatile and nonlinear data set from the original vector space by using the kernel approach. However, this is a pretty way to drag the SVR model as below:

$$y_i = \int(x_i) \sum_{i=1}^n ((\alpha_1 \alpha_m) K(x_i, x_n)) + b \tag{5}$$

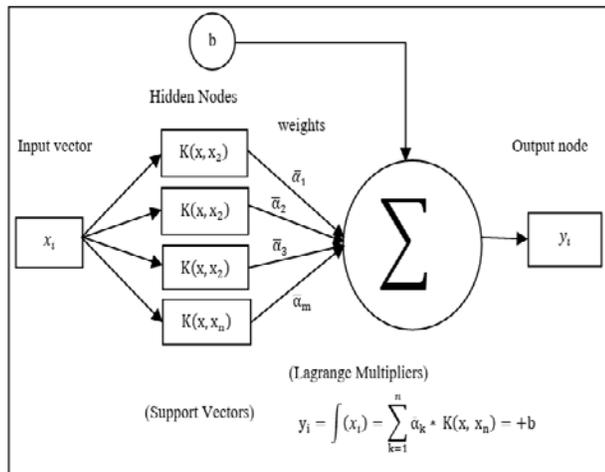


Fig. 4. Structure of Support Vector Regression.

Whereas.

Here $\alpha_1 \alpha_2$ both are represented to langrage variables multiplier. The below equation indicates that the extensive utilization of kernel by Gaussian Function concerning the size of σ^2 :

$$Z(x_i x_k) = \exp\left(-\|x_i - x_n\|^2 (2\sigma^2)\right) \tag{6}$$

2) *Random forest*: The RF is an ensemble approach which keeps the capability of performing as both regression and classification problems. In addition, it can handle binary and multi-classification problems. Moreover, the present study uses bootstrap aggregation for absolute random selection as follows [36]. For instance, when the explained variable is a continuous variable and θ is getting random vector then it promotes to decision trees of RF for regression purpose. Moreover, the x and y are random vectors and training data set taken systematically. Besides, $h_i(x)$ is a single decision tree from multi decision trees where $k - th$ denote the number of decision trees for prediction $h(x, \theta_i)$, where the $k - th$ number of outputs by $i - th$ random vectors as $i = 1, 2, 3, \dots, k$ as follows [37].

$$H(x) = \frac{1}{k} \sum_{i=1}^k h_i(x) \tag{7}$$

In equation 7, the $H(x)$ demonstrates the output of combined RF regression models. Moreover, in the training data set $k - th$ shows the number of values represented in the model.

$$H(x) = \underset{Y}{\operatorname{argmax}} \sum_{i=1}^k I(h_i(x) = Y) \tag{8}$$

Whereas $I^*(\cdot)$ is an appropriate linear function of the model and Y denotes an outcome. Similarly, $h_i(x)$ and $h_k(x)$ are training data sets that collection of X, Y vectors.

$$f(x, y) \equiv \hat{P}_k I(h_k(x) = Y) - \max_{j \neq y} \hat{P}_k I(h_k(x) = j) \tag{9}$$

For example, A is the outcome from classifier $h_k(x)$ from random vectors that can be optimum ensemble as $\hat{p}(A) =$ proportion of classifiers $h_k (1 \leq k \leq K, \text{ whenever event } A \text{ occurs. In addition, the analysts acknowledge that they optimize the model prediction performance by making more generalizations to stochastic error } (e).$

$$Pe^* = P_{x,y}(f(x, y) < 0) \tag{10}$$

According to the theorem, the stochastic problem can be determined as below equation whenever the number of decision trees may increase.

$$Pe^* \xrightarrow{K \rightarrow \infty} P_{x,y} [P_{\Theta}(h(x, \Theta) = y) \max_{j \neq y} P_{\Theta}(h(x, \Theta) = j) < 0] \tag{11}$$

In equation 11, the $P_{x,y}$ denotes problem in x, y random vectors and $k - th$ denotes the total number of decision trees in the model. In addition, if the decision trees grow up the generalization of PE will tend to upper bond. Hence, the RF algorithm has worthy convergence and can prevent over-fitting [38, 39].

3) *Machine learning-linear regression model:* The regression approach is very useful for linearity approximation. Although the traditional approaches have been employing for execution the high-frequency and nonlinear data sets, but traditional approaches (linear regression) cannot properly investigate the nonlinear data sets [12]. However, the present study is applying the ML-LRM Fig. 5 for best prediction performance. For instance, x_i is input parameter where $i - th$ is $n - th$ number of input parameters as $i = 1, \dots, 2, \dots, 3, \dots, n$ and every input parameter takes specific weight which denotes by w_i with $i = 1, 2, 3, \dots, n$ at $n - th$ number of weights. At the time of training model, the weight multiplied with input parameters and summation biases weights. Moreover, the n th denotes the number of values of the training model. Primarily, we train our model on training data set and test the model output y_i by using the testing data set and execute the prediction performance of the model.

In mapping function Φ makes typical function with x input parameter as $\Phi: X \rightarrow \mathbb{R}^N$. A linear function of hypothetical set: weight and bias are associated with mapping function as $x \rightarrow w \cdot \Phi(x) + b$: $w \in \mathbb{R}^N, b \in \mathbb{R}$. Moreover, the exclusive weight leads to risk minimization and optimization of the typical model.

$$\min_{w, b} F(w, b) = \frac{1}{m} \sum_{i=1}^m (w \cdot \Phi(x_i) + b - y_i)^2 \quad (12)$$

$$F(W) = \frac{1}{m} \|X^T W - Y\|^2 \quad (13)$$

$$X = \begin{bmatrix} \Phi(x_1) & \dots & \Phi(x_m) \\ 1 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{(N+1) \times m} \quad (14)$$

So,

$$X^T = \begin{bmatrix} \Phi(x_1)^T & 1 \\ \vdots & - \\ \Phi(x_m)^T & 1 \end{bmatrix} \quad W = \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_N \\ b \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad (15)$$

Additionally, the convex and differentiable functions executed as:

$$\nabla F(W) = \frac{2}{m} X(X^T W - Y) \quad (16)$$

$$\nabla F(W) = 0 \Leftrightarrow X(X^T W - Y) = 0 \Leftrightarrow XX^T W = XY \quad (17)$$

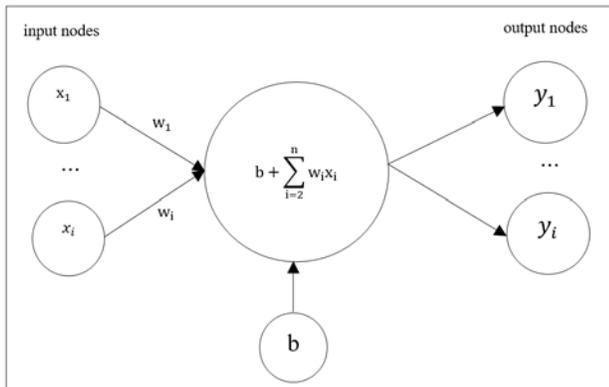


Fig. 5. Structure of ML-LRM.

Equation 17 is explained that if the $w = 0$, then the hypothesis set would also be zero / no fraction between input-output parameters of the model.

IV. RESULTS AND DISCUSSION

Table I summarizes the significant information of both group data. The group1 is contained on daily stock price returns of KSE 100-index of Pakistan and group2 is contained on PKR currency exchange rate against five major currencies of the world. The above statistics demonstrated that the total number of values are 3187 and 3464 in group1 and group2, respectively. Moreover, in group1, the mean and standard deviations are 0.04 and 1.19, respectively, indicating that the group1 data set is highly deviated over the period. In group2, the mean of currencies exchange rates is 97.06, 121.7, 146.9, 98.4 and 0.96, and the standard deviation of currencies exchange rates is 23.4, 21.3, 20.9, 26.9 and 0.22, respectively. According to group 2 statistics, the PKR has been depreciated since 2007 to 2020. Furthermore, the other descriptive statistics are illustrated in the above table as like minimum, maximum, etc.

Table II executes the augmented dickey fuller (ADF) approach to investigate whether the particular parameters have unit-root (non-stationary) or not (stationary) at specific critical level. Although the machine learning algorithms have capability to analyze the nonlinear approximation, but with respect to econometric assumptions, the empirical data should be to stationary (normal distributed) before to diagnosed. So in Table II, the statistics demonstrated that the t-value > critical value at 1%, 5% & 10% confidence interval at 1st difference. However, the unit root is not existed in the significant model.

TABLE I. DESCRIPTIVE STATISTICS

statistics	KSE 100-Index	Currency Exchange Rates				
	Price Return	PKR/USD	PKR/Euro	PKR/GBP	PKR/CHF	PKR/JPY
count	3187	3464	3464	3464	3464	3464
mean	0.04	97.0	121.7	146.9	98.4	0.96
std	1.19	23.4	21.3	20.9	26.9	0.22
min	-6.85	60.3	78.4	108.2	48.3	0.48
0.25	-0.48	83.7	111.6	132.5	78.9	0.85
0.5	0.04	98.06	118.4	141.4	103.6	0.96
0.75	0.63	104.8	131.2	159.1	109.9	1.08
max	8.60	166.7	186.5	208.4	173.9	1.54

TABLE II. AUGMENTED DICKEY-FULLER (ADF) TEST

Parameters	ADF	Critical Values		
		1%	5%	10%
Price Returns	-47.339*	-3.432	-2.8622	-2.567
PKR/USD	-52.361*	-3.432	-2.8621	-2.567
PKR/Euro	-57.412*	-3.432	-2.8621	-2.567
PKR/GBP	-56.698*	-3.432	-2.8621	-2.567
PKR/CHF	-56.854*	-3.432	-2.8621	-2.567
PKR/JPY	-57.382*	-3.432	-2.8621	-2.567

Note: * indicates no unit-root exist in the model at 1st difference

TABLE III. PREDICTION PERFORMANCE OF MLMS WITH MAE

parameters	SVR	RF	ML-LRM
price returns	0.819**	0.978	0.823
PKR-ER against five major currencies	2.718	0.1338**	6.487
	4.774	0.519**	10.967
	5.340	0.622**	9.7980
	3.044	0.424**	8.6620
	0.048	0.003**	0.1440

Note: In Table III, ** indicate comparative best prediction performance model on certain data group.

TABLE IV. PREDICTION PERFORMANCE OF MLMS WITH MSE

parameters	SVR	RF	ML-LRM
price returns	1.545**	2.036	1.559
PKR-ER against five major currencies	13.542	0.158**	73.6090
	38.536	0.570**	173.561
	48.862	0.790**	184.280
	18.531	0.423**	108.558
	0.0030	3.203**	0.02600

Note: In Table IV, ** indicate comparative best prediction performance model on certain data group.

TABLE V. PREDICTION PERFORMANCE OF MLMS WITH RMSE

parameters	SVR	RF	ML-LRM
price returns	1.243**	1.427	1.247
PKR-ER against five major currencies	3.679	0.397**	8.579
	6.207	0.755**	13.174
	6.990	0.889**	13.574
	4.304	0.650**	10.412
	0.056	0.005**	0.1690

Note: In Table V, ** indicate comparative best prediction performance model on certain data group.

Tables III, IV and V execute the prediction performance of MLMs on different nature of data sets. In 1st experiment, the present study used daily stock price returns of KSE 100-index to evaluate the prediction performances of MLMs through mean absolute error (MAE), mean Squared error (MSE) and root mean squared error (RMSE). Consequently, the output shows that the SVR gives relatively optimal prediction performance than RF and ML-LRM on group1 data set; hence, the MAE, MSE and RMSE of SVR is less than RF and ML-LRM. But the ML-LRM is giving best prediction than RF on the corresponding data group. So, on the 1st experiment, the absolute prediction performance is summarized as SVR > ML-LRM > RF. Furthermore, in 2nd experiment, we used group2 data regarding currencies exchange rates for the prediction performance of MLMs. In addition, the loss of RF < SVR < ML-LRM while the RF is giving optimal prediction performance than SVR and ML-LRM. In addition, the SVR gives best prediction than ML-LRM at the same experiment. In 2nd experiment, the prediction performance of MLMs is categorically assembled as RF > SVR > ML-LRM. In fact, the empirical data set is nonlinear nature in 2nd experiment and linear regression algorithm is operating in ML-LRM while relatively the ML-LRM doesn't properly investigate the data. On the other hand, the RF is typically used for classification and regression problems. So the RF algorithm is best recognizing the typical data set, even though the data is

nonlinear. However, the RF is comparatively given the optimal prediction performance in 2nd experiment.

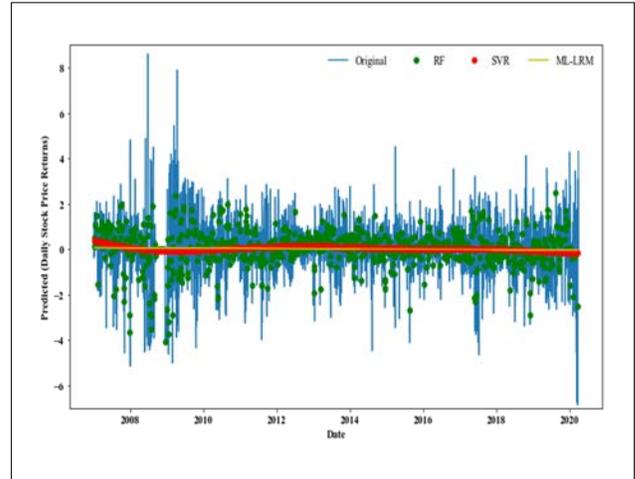


Fig. 6. Graphically Evaluation of Prediction Performance of MLMs on Group1.

Fig. 6 executes the graphical prediction performance of SVR, RF and ML-LRM at group1. The output demonstrated the SVR model is given optimal prediction than RF and ML-LRM, hence, the red line is very closed to true line. But the RF predicted plots are slightly away from true line. Similarly, the ML-LRM prediction line is comparatively less close to true line. So, the whole prediction performance of MLMs from Fig. 6 executes as follows: SVR > ML-LRM > RF. Thus, the SVR can comparatively best predict than RF and ML-LRM on a high-volatile time series data set. Some previous studies are supporting to findings of current study as: [40] determined the comparative prediction performance between SVR and ensemble empirical mode decomposition (EEMD) through Singular Spectrum Analysis (SSA) on Shanghai stock price index. Moreover, this study correspondingly examining the market tendency, market fluidity, economic topographies of the market. The output shows that the SVR can relatively best predict to the fluidity of forex market and stock market than EEMD. So, this is sound evidence about the SVR can gives more efficient and accurate prediction than generic TEM and MLMs on high-frequency data which supported to findings of current study.

The Fig. 7 executes the graphical prediction performances of five empirical models on group2. Each model has relatively summarized the prediction performance of SVR, RF and ML-LRM. In addition, the output of every model is categorically presented along with specific color as follows: SVR with dark-red, RF with yellow-green, ML-LRM with dark-orange and original input data with light-blue color. The model 7(a) in our 2nd experiment exhibits the prediction performance of MLMs by taking the PKR exchange rate against USD on the y-axis and the number of time steps on the x-axis. Consequently, the RF is giving relatively best prediction than SVR and ML-LRM while yellow green line is extremely closed to the true line. In model 7(b), again the RF is relatively giving the best prediction while the yellow green line is very closed to light blue line. Similarly, the output of models c, d & e is accompanied as previous two models. With respect to our

2nd experiment, the findings concluded the RF can comparatively best predict to the nonlinear approximation than SVR and ML-LRM. The findings of 2nd experiment is supported by [41].

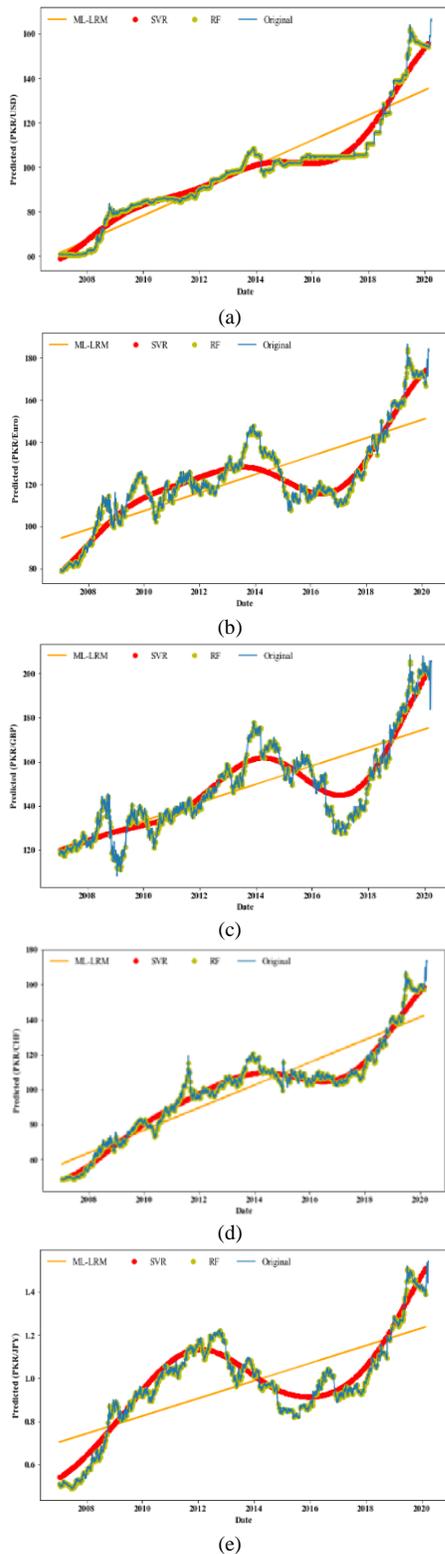


Fig. 7. Graphically Evaluation of Prediction Performance of MLMs in Group2.

V. CONCLUSION

The stock market and forex market both are highly dynamical financial markets. The usage of machine learning approaches are rapidly growing in financial markets for economical analysis. Therefore, the key purpose of the present study is to investigate the best predictor machine learning model by giving different nature of high frequency financial data sets. So, the empirical findings demonstrated that the SVR and RF have ability to optimistic prediction on high-volatile and nonlinear data sets respectively. In fact, the SVR is a supervised learning approach which works on the principle of support vector machine (SVM) to solve the classification and regression problems. Hence, the SVR is very useful approach for complex and high volatile data. On the other hand, the RF is usually used for classification problems. However, the RF has given best prediction on nonlinear approximation. Moreover, this study had targeted only forex market and stock market of a country, hence, the study may more robust by using panel financial market data. Further research directions are using hybrid deep neural networks with filter-based feature selection to improve the prediction accuracy and reduce the cost of training time and compare the performance with RF and SVR by using more complex financial data.

ACKNOWLEDGMENT

This work is supported by the National Natural Sciences Foundation of China, grant no. (71701082 and 71271103).

REFERENCES

- [1] Z.S. Zhang, Ervin Radiological images and machine learning: trends, perspectives, and prospects, *Computers in biology medicine* 108 (2019) 354-370.
- [2] A. Kuchi, M.T. Hoque, M. Abdelguerfi, M.C. Flanagan, Machine learning applications in detecting sand boils from images, *Array* 3(2019) 100012.
- [3] J.S. Almeida, P.P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, V.H.C. de Albuquerque, Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques, *Pattern Recognition Letters* 125 (2019) 55-62.
- [4] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, *Neurocomputing*, 273 (2018) 271-280.
- [5] J. Calabuig, H. Falciani, E.A. Sánchez-Pérez, Dreaming machine learning: Lipschitz extensions for reinforcement learning on financial markets, *Neurocomputing*, 398 (2020) 172-184.
- [6] B.M. Henrique, V.A. Sobreiro, H. Kimura, Literature review: Machine learning techniques applied to financial market prediction, *Expert Systems with Applications*, 124 (2019) 226-251.
- [7] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, H.E. Stanley, Which artificial intelligence algorithm better predicts the Chinese stock market?, *IEEE Access*, 6 (2018) 48625-48633.
- [8] B.J. de Almeida, R.F. Neves, N. Horta, Combining Support Vector Machine with Genetic Algorithms to optimize investments in Forex markets with high leverage, *Applied Soft Computing*, 64 (2018) 596-613.
- [9] J. Carapuço, R. Neves, N. Horta, Reinforcement learning applied to Forex trading, *Applied Soft Computing*, 73 (2018) 783-794.
- [10] A.V. Contreras, A. Llanes, A. Pérez-Bernabeu, S. Navarro, H. Pérez-Sánchez, J.J. López-Espín, J.M. Cecilia, Theory, ENMX: An elastic network model to predict the FOREX market evolution, *Simulation Modelling Practice*, 86 (2018) 1-10.

- [11] R. Chowdhury, M. Mahdy, T.N. Alam, G.D. Al Quaderi, M.A. Rahman, Predicting the stock price of frontier markets using machine learning and modified Black–Scholes Option pricing model, *Physica A: Statistical Mechanics its Applications*, 555 (2020) 124444.
- [12] H. Maqsood, I. Mehmood, M. Maqsood, M. Yasir, S. Afzal, F. Aadil, M.M. Selim, K. Muhammad, A local and global event sentiment based efficient stock exchange forecasting using deep learning, *International Journal of Information Management*, 50 (2020) 432-451.
- [13] C.-L. Huang, C.-Y. Tsai, A hybrid SOFM-SVR with a filter-based feature selection for stock market forecasting, *J Expert Systems with Applications* 36 (2009) 1529-1539.
- [14] B.M. Henrique, V.A. Sobreiro, H. Kimura, Stock price prediction using support vector regression on daily and up to the minute prices, *The Journal of finance data science* 4(2018) 183-201.
- [15] L. Guo, W. Fang, Q. Zhao, X. Wang, The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality, *J Computers Industrial Engineering* 161 (2021) 107598.
- [16] M. Vijh, D. Chandola, V.A. Tikkiwal, A. Kumar, Stock closing price prediction using machine learning techniques, *Procedia Computer Science* 167 (2020) 599-606.
- [17] P. Werawithayaset, S. Tritilanunt, Stock Closing Price Prediction Using Machine Learning, 2019 17th International Conference on ICT and Knowledge Engineering (ICT&KE), IEEE, 2019, pp. 1-8.
- [18] R. Philip, Estimating permanent price impact via machine learning, *Journal of Econometrics*, 215 (2020) 414-449.
- [19] A. Yadav, C. Jha, A. Sharan, Optimizing LSTM for time series prediction in Indian stock market, *Procedia Computer Science*, 167 (2020) 2091-2100.
- [20] Y. Zhang, G. Chu, D. Shen, The Role of Investor Attention in Predicting Stock Prices: The Long Short-term Memory Networks Perspective, *Finance Research Letters*, DOI (2020) 101484.
- [21] A. Moghar, M. Hamiche, Stock Market Prediction Using LSTM Recurrent Neural Network, *Procedia Computer Science*, 170 (2020) 1168-1173.
- [22] P. Wang, Pricing currency options with support vector regression and stochastic volatility model with jumps, *Expert Systems with Applications* 38 (2011) 1-7.
- [23] W. Chen, H. Xu, L. Jia, Y. Gao, Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants, *International Journal of Forecasting*, DOI (2020).
- [24] R. Dash, An improved shuffled frog leaping algorithm based evolutionary framework for currency exchange rate prediction, *Physica A: Statistical Mechanics and its Applications*, 486 (2017) 782-796.
- [25] Z. McGurk, US real estate inflation prediction: Exchange rates and net foreign assets, *The Quarterly Review of Economics and Finance*, 75 (2020) 53-66.
- [26] T. Lumley, Survey analysis in R, the 'survey' package). Guía de usuario disponible en: <http://faculty.washington.edu/tlumley/survey>, DOI (2010).
- [27] J. Aizenman, A. Powell, Volatility and financial intermediation, *Journal of International Money and Finance*, 22 (2003) 657-679.
- [28] P. Aghion, S.N. Durlauf, *Handbook of economic growth*, Elsevier 2005.
- [29] M. Busse, C. Hefeker, Political risk, institutions and foreign direct investment, *European journal of political economy*, 23 (2007) 397-415.
- [30] V. Vapnik, S.E. Golowich, A.J. Smola, Support vector method for function approximation, regression estimation and signal processing, *Advances in neural information processing systems*, 1997, pp. 281-287.
- [31] T. Wuest, D. Weimer, C. Irgens, K.-D. Thoben, Machine learning in manufacturing: advantages, challenges, and applications, *Production & Manufacturing Research*, 4 (2016) 23-45.
- [32] C. Cortes, V. Vapnik, Support-vector networks, *Machine learning*, 20 (1995) 273-297.
- [33] W. Fenghua, X. Jihong, H. Zhifang, G. Xu, Stock price prediction based on SSA and SVM, *Procedia Computer Science*, 31 (2014) 625-631.
- [34] D. Wang, Y. Zhao, Using News to Predict Investor Sentiment: Based on SVM Model, *Procedia Computer Science*, 174 (2020) 191-199.
- [35] P. Anand, R. Rastogi, S. Chandra, A class of new Support Vector Regression models, *Applied Soft Computing*, DOI (2020) 106446.
- [36] W. Alsuraihi, E. Al-hazmi, K. Bawazeer, H. Alghamdi, Machine Learning Algorithms for Diamond Price Prediction, *Proceedings of the 2020 2nd International Conference on Image, Video and Signal Processing*, 2020, pp. 150-154.
- [37] V. Vapnik, S.E. Golowich, A. Smola, Support vector method for function approximation, regression estimation, and signal processing, *Advances in neural information processing systems*, DOI (1997) 281-287.
- [38] S. Pan, S. Zhou, Evaluation Research of Credit Risk on P2P Lending based on Random Forest and Visual Graph Model, *Journal of Visual Communication and Image Representation*, DOI (2019) 102680.
- [39] L. Breiman, Random forests, *Machine learning*, 45 (2001) 5-32.
- [40] V. Zavgorodny, P. Lukyanov, S. Nazarov, 2nd International Conference on Information Technology and Quantitative Management, ITQM, DOI (2014).
- [41] C. Ciner, Do industry returns predict the stock market? A reprise using the random forest, *The Quarterly Review of Economics and Finance*, 72 (2019) 152-158.

Optimize and Secure Routing Protocol for Multi-hop Wireless Network

Salwa Othmen¹, Wahida Mansouri², Somia Askilany³, Wided Ben Daoud⁴
Computers and Information Technology Department, College of Science and Arts^{1,2,3}
Turaif, Northern Border University, Kingdom of Saudi Arabia^{1,2,3}
NTS'Com Research Unit, ENET'COM, University of Sfax, Sfax, Tunisia⁴

Abstract—Multi-hop Wireless Network (MWN) requires the existing of wireless nodes that communicate via a wireless channel. Thus, selecting optimal paths between the communicant nodes is a major challenge. Many researchers are focusing on this topic and proposed some routing protocols that help the nodes to learn multi-hop paths. Multi-hop wireless network is used for several types of applications, like military, medical care and national security. These applications are important and critical, so they require a certain level of performance and security during data communication. Securing the transmission of data in a multi-hop network is a challenge since the devices have limited resources like memory and battery. In this paper, we propose an optimal and secure routing protocol. The main goal of this proposal is to improve the performance and the security of such network by selecting a secure route between the source and its target destination. To secure data transmission phase, we propose to create a key shared between the source and the destination. Since the devices have limited energy, we propose to take into consideration the energy of the intermediate nodes of the selected route. Extensive simulations are performed using the Network Simulator (NS2) to validate the proposed protocol. This proposal is compared with the secured Ad-Hoc On-demand Distance Vector (SAODV) in terms of end-to-end delay, overhead and number of compromised devices.

Keywords—Multi-hop wireless network; routing protocol; Diffie-Hellman; Weil Pairing; NS2

I. INTRODUCTION

Multi-hop Wireless Networks (MWNs) like Ad Hoc networks, sensor networks and Internet of things (IoT) compose of wireless and discrete devices that communicate with each other directly without needing any fixed infrastructure [1]. Thus, the main function of such network is to route the information from the source to the destination from a node to another. This is performed by exchanging route information across different devices of the network. Many researchers are interested in this important topic and try to propose routing protocols which allow the devices to learn some multi-hop route between them. The routing protocols are classified into three categories: reactive, proactive and hybrid protocols.

In the proactive protocol, each node maintains a routing table that contains information about existing routes. When a node tries to transmit data, it uses a route already exist in this routing table. Many proactive routing protocols are proposed like Optimized Link State routing protocol (OLSR) [2], Destination Sequence Distance Vector (DSDV) [3] and

Wireless Routing Protocol (WRP) [4], etc. However, in the reactive protocol, only one active route is required to reduce the overhead in the network like Dynamic Source Routing (DSR) [5], Temporarily Ordered Routing Protocol (TORA) and Ad-Hoc On Demand Distance Vector (AODV) [6]. The hybrid protocol is a combination of the reactive and the proactive protocols.

Many challenges may affect the use of multi-hop wireless networks to support many applications due to their specific characteristics like the unstable topology, energy efficiency and mobility, etc. Therefore, it is important to take into account these challenges when designing and improving the functions of MWN such as the multi-hop routing protocols.

Another important issue must be taken into consideration when designing a routing protocol for MWN, is the security due to the participation of the nodes in the routing process. Indeed, an attacker can participate in the route discovery phase to become a member of the selected route and it can later perform different types of attacks like dropping, forging or injecting data packet. However, introducing security in multi-hop routing protocols needs extra resource consumption and extra storage due to the underlying computation cost required. This is not desirable in MWN as the limited resources of the nodes. Thus, when securing the routing in MWN, an efficient use of resources should be considered, in particular the energy resource.

However, many proposed protocols handle the resource efficiency and the security separately.

The security of routing protocols is mandatory to provide the protection of the exchanged data from the source node to the destination node. Most of the proposed protocols ensure the security from sharing secret keys between each two neighbor nodes. Thus, the number of the shared keys increases with the increase of the nodes in the network. This is lead to high resource consumption, which is not efficient, especially for some critical networks like the sensor networks due to the limited resources of the sensor nodes.

In this paper, we propose a new multi-hop routing protocol for MWN. This proposal selects secure and optimal path that ensures security in terms of authentication, confidentiality and integrity. For achieving the anonymity, we propose to use a temporary identity for each node in the communication process. The battery life of the selected nodes is taken into account in the proposal to achieve the performance of such

network. Indeed, during the route discovery process, only the nodes with high energy can be selected. When a destination receives several request packets, it selects the shortest and the longest lifetime route based on a proposed cost function.

The current paper is organized as follows: Section 2 offers an overview on some related works. Section 3 provides a detailed description of the proposed routing protocol. The last section makes the conclusion of the work and suggests the future research.

II. RELATED WORK

Many routing protocols are proposed in the literature to optimize the performance of the MWNs. Indeed, in [7], a multi-hop routing protocol using cellular virtual grid in internet of thing environment is proposed. The goal of this proposal is to prolong the network lifetime through the balancing of energy consumption. The cost of the path between the source and the destination nodes is computed based on the residual energy and the distance. In [8], O. Salwa et al. proposed a fuzzy logic based on-demand routing protocol for multi-hop cellular networks. To optimize the performance of the network, the authors combine three metrics which are Signal to Interference and Noise Ratio (SINR), residual energy and gain time, based on the fuzzy logic system. In [9], a low-overhead multi-hop routing protocol for device to device communication in 5G is proposed. The proposal is based on the DSR protocol to select an optimal route for 5G in a short time. The overhead is reduced through the minimizing of the exchanged control messages, so the time and the energy are saved during the route discovery process. These three protocols [7-9] improve the networks performance in terms of energy consumption and lifetime.

However, the main drawback of these proposals is that they do not take into account the security requirements.

Other works are proposed to secure and optimize the routing protocol for multi-hop networks. Indeed, in [10] H. Kojima et al. proposed to secure DSR protocol using sequential aggregate signature in order to sign the routing information. In this proposal, the communication between devices requires a centralized key generation center to distribute the keys in the network. Thus, any new device cannot join the network without authentication to this center. In [11], G. Singh et al. proposed a routing protocol called Expiration Time based Routing Protocol (LETSRP) which based on a one-time signature scheme to authenticate the exchanged data in the network. Before sending any packet, each node computes the time expiration of its links using a greedy algorithm. The number of the sent packets depends on the available bandwidth. In [12], a secured and optimized routing protocol for MANET is proposed by A. Bhusari et al. This work was designed to optimize the performance of the routing protocol by minimizing the overhead and the delay. To secure this proposed protocol, a new metric based on cross layer design is provided to defend several attacks. To secure the request phase, the source node signs the RREQ with the group signature based on its private key. The destination node decrypts the received packet using the public group signature key.

However, the disclosure of the generated signature may cause the disclosure of the entire network. This is because the nodes use the same key during the communication process.

In [13], A. Vinitha et al. proposed a secure multi-hop routing protocol for wireless sensor networks. To secure the proposed protocol, the authors employed a trust model using several trust factors like indirect trust, direct trust, forward rate factors and integrating factor. To ensure the optimization of the proposal, the trust factors are integrated with other parameters such as delay, distance, energy, intra-cluster distance and inter-cluster distance. Before selecting the optimal route, the network is devised into a several cluster. The cluster heads are selected based on the Low Energy Adaptive Clustering Hierarchy (LEACH) protocol.

In [14], K. Hamouid et al. proposed a secure tree-based routing protocol for wireless sensor networks. The authors use ID-Based authentication key-agreement protocol to secure the data routing between the nodes in the network. The confidentiality and the authenticity are provided in the proposed protocol with low cost. Indeed, each node in the network is preloaded with a private key used to generate shared keys with its neighbor nodes. Moreover, to reduce the communication overheads a single message is transmitted by each node for both key establishment and routing-tree construction. However, in this protocol, each node must perform complex operations to generate security keys. This may increase the energy consumption by the nodes.

In [15], Zapata et al. proposed a secure routing protocol called Secure Ad Hoc On-Demand Distance Vector (SAODV) which is an extension of the AODV protocol to guarantee its security in terms of authentication, integrity and non-repudiation. To achieve the authentication, the source and the destination nodes add their signatures based on their private keys. The intermediate nodes only check the validity of this generated signature without any authentication performed between each other. However, to achieve the integrity of the hop-count field, a hash chain is used. The function used to compute the hash value is added to the hash function field. The SAODV protocol uses several mechanisms to secure the route request phase, but it remains vulnerable to many types of attacks. This fact is due to the lack of the authentication between neighbor nodes. Indeed, an adversary can participate in the selected path without modifying the hop-count field by using the same hash value. Thus, the legitimate nodes cannot detect this attack. In [16], M. Surajuddin et al. proposed a routing protocol that takes into account multiple factors such as packet loss reduction, congestion, malicious node detection and security of data transmission. Indeed, the source broadcasts a RREQ packet, which contains a fake destination address and sequence number. Only an attacker will respond with a RREP packet. In this case, the source maintains the address of this attacker in a black list and propagates this information to the other nodes in the network. Moreover, each node has a trust value calculated based on the opinion of its neighbors. Through this trust value, the nodes can identify the malicious nodes which have a trust value less than a threshold. However, this proposed protocol is not secured against several types of attacks like the impersonation attack and Sybil attack.

To overcome some limitations of the existing works like the lack of authentication between the neighbor nodes and the complex operations to compute a shared key, etc., we propose a new protocol described in the following section.

III. PROPOSED PROTOCOL

A. Weil Pairing

In the proposed protocol, we are based on Weil Pairing tool for key generation. Indeed, the Weil Pairing [17] is an important method used in elliptic curve systems, key generation and identity-based encryption.

Let two groups G_1 and G_2 of order q , note that G_1 is an additive cyclic group over an elliptic curve and G_2 is a multiplicative cyclic group. P is a generator of G_1 . The admissible bilinear map:

$\hat{e}: G_1 \times G_1 \rightarrow G_2$ has the following properties:

- Bilinear: for all $P, Q \in G_1$ and for $a, b \in \mathbb{Z}$, $\hat{e}(aP, bQ) = \hat{e}(P, Q)^{ab}$.
- Non-degenerate: if $P, Q \in G_1$ such that $\hat{e}(P, Q) \neq 1$.
- Computable: for all $P, Q \in G_1$, $\hat{e}(P, Q)$ can be computed efficiently.

B. Network Model

In the proposal, we consider a MWN which consists of multiple devices distributed randomly in a geographical area and a trusted party (TP). Each device has a unique identity in the network (ID). We assume that these devices are not secured and so they can be compromised, but we suppose that TP is secure and trustworthy. Thus, we suppose that the TP is responsible for the generation and the record of the system parameters in a secure way.

These parameters are as follows:

- p is a large prime number,
- G_1 and G_2 are two cyclic groups.
- g is a generator in \mathbb{Z}_p^* ,
- P is a generator of G_1 ,
- H : is a hash function, $\{0, 1\}^* \rightarrow G_1$,

The TP generates a private key called $s \in \mathbb{Z}_q^*$, and a master key $P_{pub} = sP$. Then, it bootstraps the devices with initial secret parameters in offline before network deployment. Indeed, it assigns a private key for each device; $S_i = sQ_i$, where $Q_i = H(ID_i || t)$, t is the timestamp initiated by the TP in order to prevent the network against replay attack. Before the route discovery process, each device D_i must compute and share a secret key with the devices located at n hops. For that reason, it generates a random value called R_i and sends to the neighbor devices the following value: $P_i = R_iP$ used to compute the shared key. This value is based on Pairing Discrete Logarithm Problem (PDLP) which is a complex problem because finding the integer R_i is hard.

C. Description of the Proposed Algorithm

The proposed routing protocol is divided into three phases: route request, route reply and data transmission phases. This protocol is proposed for MWN which is a hostile environment where it can be intercepted by different types of attacks. Thus, a high level of security must be achieved to secure each phase among the phases listed below. In addition to security challenge, the MWN face other major challenges like energy constraint of devices. Indeed, energy is a critical resource in MWN as its lifetime depends on battery depletion of mobile devices. More energy consumed in the routing process leads to reduce the network lifetime. For that reason, in the proposed routing protocol, we ensure an efficient use of the limited resources, in particular the energy consumption.

1) *Route request phase*: When a source device (S) intends to communicate to a destination device (D) and it does not have a valid route, it initiates the route request phase by broadcasting a route request (RREQ) packet to all its neighbors. It is based on Location based-Multiple metric (LoMM) to reduce the number of the devices that can receive the RREQ packet. This is to exclude the devices that are further away from D to participate in routing data. By this way, the end-to-end delay and the signaling load are reduced. Moreover, this metric reduce the complexity of the computational operations as all devices are participating in computing a group key and so reducing the energy consumption in the network.

To secure the request phase, each two neighbor devices share a secret key. During this step, the two devices perform a mutual authentication between each other at the same time. To reduce the computational complexity, the neighbors perform just a single evaluation of the Weil Pairing as compared with other schemes like Smart-Chen-Kudla scheme.

The generation of the shared key K_{ij} between each two neighbor devices D_i and D_j is performed as follows:

D_i computes K_{ij} using the following equation (1):

$$\begin{aligned} K_{ij} &= \hat{e}(S_i; R_jQ_j + R_iQ_j) \\ &= \hat{e}(sQ_i; R_jQ_j + R_iQ_j) \\ &= \hat{e}(Q_i; Q_j)^{s(R_i+R_j)} \end{aligned}$$

In the other side D_j computes also K_{ji} as the following function:

$$\begin{aligned} K_{ji} &= \hat{e}(R_iQ_i + R_jQ_j; R_j) \\ &= \hat{e}(R_iQ_i + R_jQ_j; sQ_j) \\ &= \hat{e}(Q_i; Q_j)^{s(R_i+R_j)} \\ &= K_{ij} \end{aligned}$$

Where, R_i and R_j are random values generated by D_i and D_j respectively, and exchanged based on Pairing Discrete Logarithm Problem as mentioned above.

The source initiates the route request phase by broadcasting a RREQ packet to all its neighbors. The format of the RREQ is as follows:

RREQ: {ID_S, E(K_{sj}, ID_D || seqNb || TTL || Hop-count || K_{PK_S}(g^{R_S mod p) || ME) || MAC_{Ksj}(ID_S, ID_D, seqNb, TTL, Hop-count, K_{PK_S}(g^{R_S mod p), ME))}}}

Where,

- K_{sj} is the shared key between S and each of its neighbor D_j. It is calculated as the function (1) in order to encrypt the RREQ packet between each other and so to provide the confidentiality of this packet.
- ID_S is the source address,
- ID_D is the destination address,
- seqNb is the sequence number which prevents the RREQ packet against replay attack,
- TTL: is the Time To Live which limits the propagation area of the RREQ packet,
- Hop_count: is a value incremented by each intermediate node to count the number of hops in the discovered route,
- g^{R_S mod p}: is a value used to compute a shared key between S and D, where R_S is a random number generated by S. For more security, this value is encrypted by the private key of the source. Finding R_S is hard as it is difficult to resolve the Diffie-Hellman problem in prime order.
- ME: is the minimum remaining energy which represents the lifetime of the discovered route.
- MAC_{Ksj}: is a function used to check the integrity of the RREQ packet.

When an intermediate device D_i receives a RREQ packet from a neighbor D_j, it performs the following steps:

- Decrypts the received packet using the shared key K_{ij} with the sender device, which is calculated as the function (1). If this decryption is performed successfully, so that a mutual authentication is provided between them because only these two devices can calculate this shared key.
- Computes the MAC function to verify the integrity of the received packet. If there is no problem with the integrity, D_i passes to the next step, otherwise, it discards the received packet.
- Checks if it is the target destination by comparing its own address and ID_D. If it is the target destination, it sends back a response to the source via the reverse route if not, it performs the next steps,
- Checks if TTL is zero, it discards the received packet, if not it decrements this field and increments hop_count field,

- Computes its residual energy and compares it with the ME field, then it reassigns the ME field with the minimum value among them.
- Maintains the address of the sender in its routing table, and adds its address in the RREQ packet.
- Computes the MAC function using the key shared with each neighbor,
- Sends the RREQ packet to the neighbors after encrypting it by the secret key shared with each of these neighbors.

The RREQ packet is sent until it reaches the destination in a secure way.

D. Route Reply Phase

When the destination receives a RREQ packet, it waits for a definite time to receive other RREQ packets. Then it performs the following steps:

- It decrypts the received packets and checks their integrity. Then, it calculates the cost (C) of each discovered path as the following equation:

$$C = \frac{ME}{hop_count}$$

- It selects the path with the largest cost C. By this way, it chooses the path that has the greatest remaining energy and the smallest number of intermediate nodes (shortest path).
- Computes the shared key with the source based on Diffie_Hellman problem as the following equation:

$$K_{DS} = g^{R_S * R_D} \text{ mod } p$$

Where, R_D is a random value generated by the destination. R_D is sent to the source using Diffie_Hellman problem $g^{R_D} \text{ mod } p$ to recalculate the shared key with the source as follows:

$$K_{SD} = g^{R_D * R_S} \text{ mod } p$$

Thus, the same key is obtained by the source and the destination:

$$K_{SD} = K_{DS}$$

- Generates the RREP packet:

$$\text{RREP: } \{ID_D, E(K_{Dj}, ID_S || ID_D || K_{PK_D}(g^{R_D} \text{ mod } p) || MAC_{K_{Dj}}(ID_S, ID_D, K_{PK_D}(g^{R_D} \text{ mod } p)))\}$$

The RREP is encrypted by K_{Dj} which is the shared key between the destination and the intermediate device j of the selected route. K_{Dj} ensure also a mutual authentication between the two communicants.

MAC_{K_{Dj}} is used to check the integrity of the packet.

$g^{R_D} \text{ mod } p$ is encrypted by the private key of the destination for more security.

The RREP is sent through the reverse route until it reaches the source device.

When the source receives the RREP packet, it re-computes the shared key with the destination and triggers the transmission data phase secured by the shared key between the source and the destination.

IV. SECURITY ANALYSIS

In the following section, we analyze the security of the proposal against several threats by showing that it achieves the security constraints:

A. Confidentiality

The messages exchanged in the request and reply phases of the proposed protocol are encrypted with the keys shared between the neighbor devices. To compromise these keys, the attackers need to know the secret parameters used to calculate each key, and so they have to resolve the PDLP which is a hard problem.

Furthermore, the data transmission phase is secured based on the shared key between the source and the destination. To compromise this key, the attacker needs to resolve the Diffie-Hellman problem.

Thus, the proposed protocol ensures the confidentiality of the exchanged messages.

B. Authentication

In the proposed protocol, each two neighbors have to share a secret key based on Weil Pairing scheme. This method provides an implicit mutual authentication between the communicants using some secret parameters. Indeed, every device computes and shares a secret key with its neighbor; only a legitimate device can compute this key as it is based on the private key of the TP. Moreover, the source and destination authenticates each other through the shared key between them, which is calculated based on Diffie-Hellman problem.

Thus, the proposed protocol achieves the authentication.

C. Integrity

In the proposal, to check the integrity of each transmitted message, the sender adds the MAC function to this message. To forge the integrity of such packet, an attacker must decrypt it and re-compute the MAC function of the modified packet. However, this is not possible as the attacker does not learn the secret key used to encrypt the received packet.

D. Sybil Attack

Sybil attack occurs when an attacker use unauthorized identities to perform neighbor relationships with other legitimates devices. In the proposal, when an attacker sends messages to a legitimate device using a forged identity, it fails in performing a mutual authentication as it has not a valid key issued by the TP. Thus, to perform a Sybil attack, the attacker must generate its own private key, which is impossible because it is hard to solve PDLP problem and hold the private key of the trust party. For that reason, the proposal is secured against Sybil attack.

E. Replay Attack

The attacker tries to falsify the destination by retransmitting many authorized packets. The proposed protocol is secured against this type of attack for many reasons. First, because the source generates a sequence number for each new request packet. Second, the private key of each device is computed based on a timestamp initiated by the source. Moreover, the shared keys are based on a random number generated by legitimates devices in each session without any links with the values generated in the previous session. Thus, the proposed protocol is secured against replay attack.

F. Impersonation Attack

In this type of attack, the attacker uses a legitimate identity to perform a neighbor relationship or to participate in the selected route as an intermediate device. In this proposal, to impersonate a device, the attacker must compute a shared key with this device. However, it cannot obtain the same key computed by the legitimate device as it does not hold a private key assigned by the TP. Then, it is not feasible to resolve the PDLP and discover the private key of TP. Thus, it is not possible to impersonate a legitimate device.

V. SIMULATION RESULT

To evaluate the performance of the proposed routing protocol, we conduct extensive simulations using the network simulator (NS-2). We add to this simulator the security library Crypto++ as it supports many tools of security mechanism. The network is composed of 60 devices that move by Two Ray Ground model in 1000m*1000m area.

The parameters of the simulation are summarized as the following Table I.

The compromise of the legitimate devices is a major challenge which is hard to defend. If a device is compromised, the attacker can participate in the selected route and access to the exchanged messages and security parameters. Thus, all the devices can be affected. To evaluate the robustness of the proposed protocol against the malicious nodes, we introduce several attackers that held black hole attack. They pretend to be the target destination by sending RREP packets while it receives a RREQ packet, or they try to become members of the selected route.

TABLE I. SIMULATION PARAMETERS

Parameters	Value
Routing protocols	Proposed protocol, SAODV
Simulation time	200 seconds
Simulation area	1000*1000
Traffic type	Constant Bit Rate (CBR)
Packet size	512 bytes
Queue length	250 packets
MAC protocol	MAC/802.11
Mobility model	Two Ray ground
Initial energy	150J
Transmission energy	0.5 W

The proposed protocol is compared with the SAODV protocol, then, we measure three metrics as follows:

- 1) End-to-end delay: is the average delay between the time of packet generation and the time of its reception by the receiver.
- 2) Overhead: is the average of the amount received messages by each device during the route establishment phase.
- 3) Number of compromised devices: is the total number of devices compromised by the attackers during simulation time.

Fig. 1 represents the results of the end-to-end delay as a function of the number of attackers. As we can see, the proposed protocol has less value of end-to-end delay than SAODV protocol. This is due to the fact that, the proposal selects the shortest secured path. Moreover, in the proposal, the risk that an attacker compromises a device and participates in the selected route is less than SAODV protocol. Indeed, if an attacker becomes a member of the selected route, it maintains the received packets more time to handle its content and extracts the needed information from these packets. Therefore, the delay required for a packet to reach the destination is increased.

Fig. 2 presents the results of the overhead versus the number of attackers. As we can see, in the proposed protocol the message load is reduced compared with the SAODV protocol when the number of attackers increases. This is because; in the proposal when a device receives a packet from an attacker it drops this packet, but with SAODV the devices resent every received packet. Indeed, in our proposal, the neighbor devices authenticate each other by checking the shared key used to encrypt the received packet. However, in SAODV no mutual authentication is achieved.

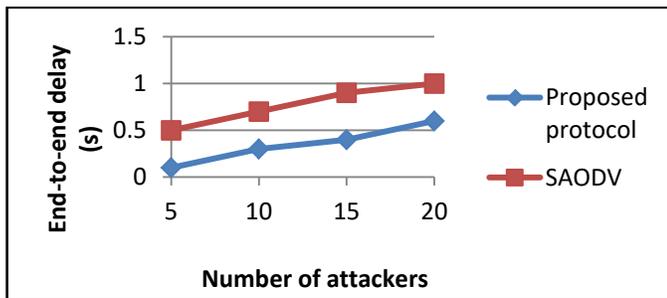


Fig. 1. End-to-end Delay Versus Numbers of Attackers.

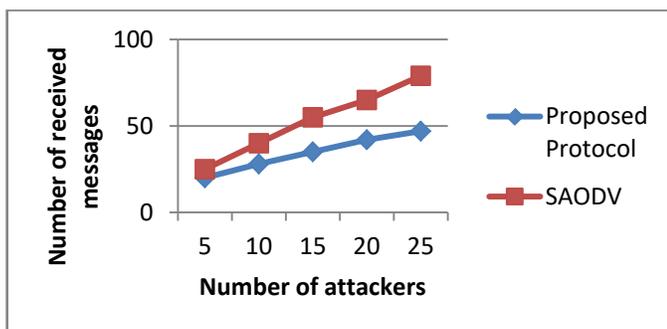


Fig. 2. Overhead Versus Number of Attackers.

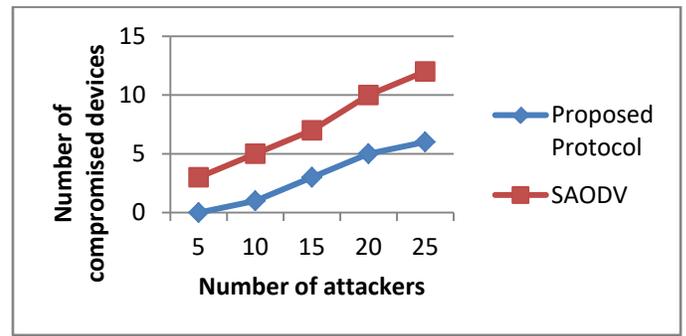


Fig. 3. Number of Compromised Devices Versus Number of Attackers.

Fig. 3 shows the results of the number of compromised devices as a function of the number of attackers. As we can see, the number of compromised devices in our proposed protocol is very less than SAODV protocol. This is because in our proposal when a device is compromised, only its private key is also compromised. The other devices are not affected, as the attacker cannot perform a mutual authentication with them because it cannot compute shared keys with these legitimate devices. However, in SAODV protocol from a compromised device, an attacker can compromise also its neighbor devices as the mutual authentication is achieved only between the source and the destination.

VI. CONCLUSION AND FUTURE WORK

The special characteristics of the MWNs have a major impact on the security of routing data between the communicants. Indeed, secure a routing protocol in this type of network is exposed to many challenges like the limited battery of the devices and their small memories. Moreover, the routing is performed hop by hop through ordinary nodes, so an attacker can easily compromise some nodes and participate in the selected route. In this context, we have proposed a secure and optimal routing protocol for MWN. This proposal takes into consideration the battery life of the intermediate devices that participate in the selected route. Moreover, security requirements like the confidentiality and authenticity are achieved during the routing process based on a proposed key-agreement method. Integrity is also achieved through the verification of the MAC function. To secure the request phase, we assumed that the neighbor devices compute shared keys between each other based on Weil Pairing scheme. To secure the data transmission phase, the source and the destination share a secret key where the parameters exchanged during the request phase. In this proposed protocol, we tried to fit inexpensive cryptography mechanisms in each phase to make it robust against many types of attacks.

As a future work, we plan to integrate an intrusion detection system to detect the malicious nodes and so to improve more the security in MWNs.

ACKNOWLEDGMENT

The author gratefully acknowledge the approval and the support of this research study by the grant no SAT-2018-3-9-F-7704 from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A.

REFERENCES

- [1] S. Corson and J. Macker, "Mobile Ad Hoc Networking (MANET): Routing Protocol Performance Issues and Evaluation Consideration", <http://www.ietf.org/rfc/rfc2501.txt>.
- [2] T. Clausen and P. Jacquet, "Optimized Link State Routing Protocol (OLSR)", RFC 3626, October 2003.
- [3] C. Perkins and P. Bhagwat, "Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for Mobile Computers", *Computer Communications Review*, pp.234-244, October 1994.
- [4] B. Liang and Z. J. Haas, "Hybrid Routing in Ad Hoc Networks with a Dynamic Virtual Backbone", *IEEE Transactions on Wireless Communications*, vol. 5, No. 6, pp. 1-14, June 2006.
- [5] D. B. Johnson and D. A. Maltz, "Dynamic Source Routing in Ad-Hoc Wireless Networking", in *Mobile Computing*, T. Imielinski and H. Korth, editors, Kluwer Academic Publishing, 1996.
- [6] C. Perkins, E. M. Royer and S. R. Das, "Ad Hoc On-Demand Distance Vector Routing (AODV)", RFC 3561, July 2003.
- [7] H. Zhand, "A WSN Clustering Multi-Hop Routing Protocol Using Cellular Virtual Grid in IoT Environment", *Mathematical Problems in Engineering*, vol. 2020, pp. 1-7, 2020.
- [8] S. Othmen, S. Asklany, M. Wahida, " Fuzzy Logic Based On-demand Routing Protocol for Multi-hop Cellular Networks (5G)", *IJCSNS International Journal of Computer Science and Network Security*, VOL.19 No.12, December 2019.
- [9] R. E. Ahmed, " A Low-Overhead Multi-Hop Routing Protocol for D2D Communications in 5G", *Journal of Communications Vol. 16, No. 5, May 2021*.
- [10] H. Khojima, N. Yanai and J. P. Cruz, " Improving the Security and Availability of Secure Routing Protocol", *IEEE Access*, Vol. 7, pp. 1-20, May,13, 2019.
- [11] G. Singh, H. Rohil, R. Rishi and V. Ranga, "International Journal of Engineering and Advanced Technology (IJEAT)", Vol. 9, pp. 498-504, October, 2019.
- [12] A. Bhusari, P.M. Jawandhiya and V.M.Thakare, "Optimizing performance of Anonymity based Secure Routing Protocol utilizing Cross layer Design for Mobile Adhoc Networks", *Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, pp. 1-6, 2018.
- [13] A. Vinitha, M.S.S. Rukmini and Dhirajsunehra, "Secure and energy aware multi-hop routing protocol in WSN using Taylor-based hybrid optimization algorithm", *Journal of King Saud University Computer and Information Sciences*, vol. 2019, pp. 1-12, 2019.
- [14] Khaled Hamouid, Salwa Othmen and Amine Barkat, "LSTR: Lightweight and Secure Tree-Based Routing for Wireless Sensor Networks", *Wireless Personal Communications*, vol. 2020, pp. 1-22, 22 January 2020.
- [15] M. G. Zapata, "Secure Ad hoc On-Demand Distance Vector (SAODV) Routing", *ACM SIGMOBILE Mobile Computing and Communications Review*, vol 6, pp. 06-107, July 2005.
- [16] M. Sirajuddin, Ch. Rupa, C. Lwendi and C. Biamba, "TBSMR: A Trust-Based Secure Multipath Routing Protocol for Enhancing the QoS of the Mobile Ad Hoc Network", *Security and Communication Networks*, vol 2021, pp. 1-9, April 2021.
- [17] Boneh, D and Franklin, M. K., " Identity-based encryption from the Weil pairing. In 21st annual international cryptology conference advances in cryptology—CRYPTO 2001, Santa Barbara, California, USA, August 19–23, Proceedings, pp. 213–229, 2001.

A Machine Learning Approach to Weather Prediction in Wireless Sensor Networks

Mrs Suvarna S Patil¹
Assistant Professor
Department of E&CE
RYMEC, Ballari

Dr B.M.Vidyavathi²
Professor and Head
Department of Artificial Intelligence and Machine Learning
BITM, Ballari, India

Abstract—Weather prediction is the key requirement to save many lives from environmental disasters like landslides, earthquake, flood, forest fire, tsunami etc. Disaster monitoring and issuing forewarning to people, living in disaster-prone places, can help protect lives. In this paper, the Multiple Linear Regression (MLR) model is proposed for humidity prediction. After exploratory data analysis and outlier treatment, Multiple Linear Regression technique was applied to predict humidity. Intel lab dataset, collected by deploying 54 sensors, to form a wireless sensor network, an advanced networking technology that existed in the frontier of computer networks, is used for solution build. Inputs to the model are various meteorological variables, for predicting weather precisely. The model is evaluated using metrics - Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). From experimentation, the applied method generated results with a minimum error of 11%, hence the model is statistically significant and predictions more reliable than other methods.

Keywords—Data mining; wireless sensor network; multiple linear regression; outliers treatment; r-square; adjusted r-square

I. INTRODUCTION

Earlier information processing was done using general purpose devices like Mainframes, laptops, palmtops etc. In many applications these computational devices are used to process human centered information. But in some applications, controlling and monitoring action is required by focusing on physical environment. For example, in a chemical factory, processes can be controlled for exact temperature. Here controlling operation is embedded with computation without human intervention. Due to the technological advancement, another important aspect needed along with computation and control is communication. This processed information needs to be transferred to the place where it is necessary, a user or an actuator. Wired communication is expensive compared to wireless communication; even wires restrict devices from moving and prevent sensors and actuators being close to the event under observation. Hence, an implementation of a new network called wireless sensor network appeared. Sensor networks are built with a number of sensors, having sensing, processing and communicating capabilities, used in real time data analysis and monitoring applications like habitat monitoring, healthcare applications, environmental monitoring and tracking of objects to mention a few. Nodes are not costly but have memory, processing and energy limitations. An example sensor network is as shown in Fig. 1.

Recently Sensor Networks (WSN) have transformed largely due to the advancement in wireless communications, Micro Electro Mechanical Systems (MEMS), distributed processing and embedded systems. These networks are widely used in various areas such as agriculture monitoring, monitoring of habitat and surveillance [1]. The most crucial system of real time monitoring and controlling is environment. Nodes in WSN are small in size and consist of microcontroller, transceiver and memory, capable of short range communication. These battery operated nodes measure temperature, humidity, light and voltage of natural event from the place of deployment and send to the sink node. Sink nodes with enough processing capability compared to end nodes perform required pre-processing on the received raw data of sensors and forward to the base station. Base stations with embedded controlling and monitoring functionalities further process the data collected from sink node for knowledge extraction and decision making to ultimate user.

Weather forecasting is a major challenge in the meteorology department due to recurrent climatic changes [2]. There are very few solutions in generating weather reports with several limitations [3]. Many outdoor activities are affected due to wind chill, rainfall and snow, the results of frequent changes in weather [4]. Inaccurate weather reports will put someone into a dangerous state [2] if the climatic conditions are not safe. There are many existing data mining techniques for processing and evaluating huge amount of weather dataset. To predict weather, data mining process has three stages—data pre-processing, Model training and then prediction.

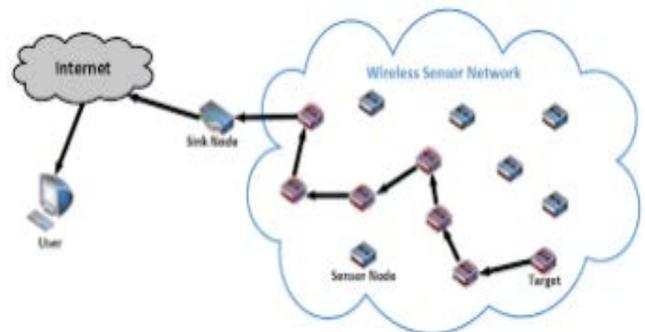


Fig. 1. An Example Wireless Sensor Network.

II. RELATED WORK

An important section of research work is literature survey, facilitating researcher to gain knowledge in the relevant field and identifying challenges. This section presents the identified best works of considered domain.

N. Krishnaveni and A. Padma [5] introduced a decision tree based algorithm called SPRINT, which builds and constructs a decision tree with relevant data. Here an enhancement in observation is made using an open source historical dataset collected from weka tool (<https://storm.cis.fordham.edu/~gweiss/datamining/datasets.html>), a tool which permits direct mining of SQL databases. Using weka on the weather parameters of considered dataset, like temperature, outlook, and humidity and windy, weather is predicted as sunny, rainy or overcast. Results proved that SPRINT is more efficient and precise compared to the existing Navie Bayes algorithm proving the performance of the work.

Munmun et al. [2] proposed an integrated method for predicting weather in order to analyze and measure environmental data. Classification is done using Naive Bayes and Chi-square algorithms. A web application states weather information taking inputs as temperature, current outlook, wind and humidity condition. Accordingly implemented system is capable of predicting weather.

Taksande et al. [6] presented forecasting of weather by Frequent Pattern Growth Algorithm. Predicting rainfall is the major goal of implementation. Dataset was collected by Nagpur station from Jan 2010–Jan 2014 and computed using Frequent Pattern Growth Algorithm. Defined variables used for predicting rain are temperature, humidity and wind speed. The implemented model worked on these parameters and provided 90% of attainment. Wang et al. [7] implemented data mining method using cloud computing in order to predict weather. Decision tree and Artificial Neural Network Algorithms are applied on meteorological data gathered at appropriate time and place. This method worked effectively on averaged weather parameters and resulted in better classification.

Sushmitha Kothapalli et al. [3] presented Auto-Regressive Integrated Moving Average (ARIMA) model for analysing and forecasting real time data. The dataset contained humidity, temperature, wind and rainfall as variables. This gathered data was stored as CSV, JSON, and XML formats in the cloud. In this work, the author was able to achieve efficient results with correlation analysis on values followed by ARIMA model.

Salman et al [8] developed a deep learning method for predicting weather. The idea was to explore internal hierarchical pattern of the dataset. For experimenting, BMKG (Indonesian Agency for Meteorology, Climatology, and Geophysics) data was considered. Recurrence Neural Network (RNN), Conditional Restricted Boltzmann Machine (CRBM) and Convolutional Network (CN) models were used on the dataset. Prediction results of these models helped the agriculture and tourism sectors.

Maqsood et al [9] introduced a group of neural networks to predict weather. Here, the groups of artificial neural networks (ANNs) were compared considering relative humidity, wind speed and temperature as the key parameters. The predictive

models used for experimenting were radial basis function network (RBFN), Elman recurrent neural network (ERNN), Hopfield model (HFM), multi-layered perceptron network (MLPN) and regression techniques. Comparative analysis showed RBFN model is better while HFM gave lesser accuracy.

Almgren et al [4] utilized Hadoop distributed system for climatic prediction. This work showed that, prior prediction diminishes event planning disasters. Here data is stored in HDFS and then processed by MapReduce programming. Outdoor events can then be planned, by obtaining processed results about weather, location and time. Oury and Singh et al. [10] created Hadoop technique for weather data analysis. Climatic conditions were investigated using precipitation, snowfall and temperature as evaluation parameters. Utilizing Apache PIG and Hadoop map reduction executed dataset. Python language was used to present output in visual form.

Manogaran and Lopez et al. [11] introduced spatial cumulative sum algorithm to detect climatic changes. MapReduce technique was applied on weather data stored in Hadoop Distributed File System (HDFS). Climatic changes were detected by applying spatial autocorrelation. Mahmood et al. [12] produced a data mining technique for predicting weather. This paper presents a data mining technique called Naive Bayes algorithm.

III. PROPOSED WORK

The newly-created model considers meteorological data to predict values for humidity by technically analysing the data and then applying multiple linear regression algorithms. In the previous work of data cleansing and pre-processing step, it was found that variables - humidity, light and voltage had a few missing values, but since the percentage of missing values is very negligible (<2%), can omit them. In the initial phase of technical analysis, data pre-processing is carried-out to gain insights into underlying source data. Exploratory analysis was carried out on pre-processed data to understand underlying relationships between dependent and independent variables. In the next phase multiple linear regression algorithms is applied on the processed data, considering key attributes as voltage, light, temperature and humidity. The model is built, trained and validated by dividing the data into Train and Test sets. For experimenting, there are many ways to partition data. But the most approved one is partitioning data into Train/Test sets or cross validation. The first set called Train data is used for model fitting and the second set called Test data to test trained model. This was followed by Model build, and, later test data was used to score the Model and obtain predicted value, which is then validated.

The master dataset needs to be divided into training and testing data - 70 percent is trained and 30 percent is testing data. Previous works presented exploratory analysis on pre-processed data to understand underlying relationships between dependent and independent values. This was followed by Model build, and, later testing data was used to score the Model and obtain predicted value, which is then validated. In the initial phase of data [13] analysis having 0.2 million samples, outliers treatment is carried out to handle extreme values.

The proposed method is shown in Fig. 2.

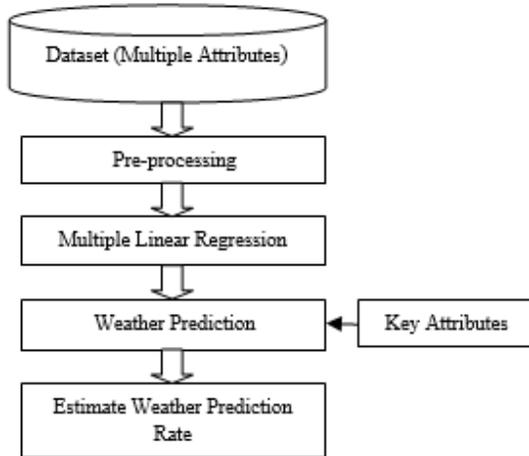


Fig. 2. Proposed Method Block Diagram.

A. Outliers Treatment

A measurement of variability called interquartile range (IQR) can be obtained by dividing data into quartiles. Depending on division the values are named as first, second, and third quartiles; denoted with Q1, Q2, and Q3, respectively. In the initial set of data, Q1 is the “middle” value given by equation 1.

$$Q1 = \left(\frac{n+1}{4}\right)^{th} \text{ term} \quad (1)$$

Median is Q2. Middle value is Q3 given by equation 2.

$$Q3 = \left(\frac{3(n+1)}{4}\right)^{th} \text{ term} \quad (2)$$

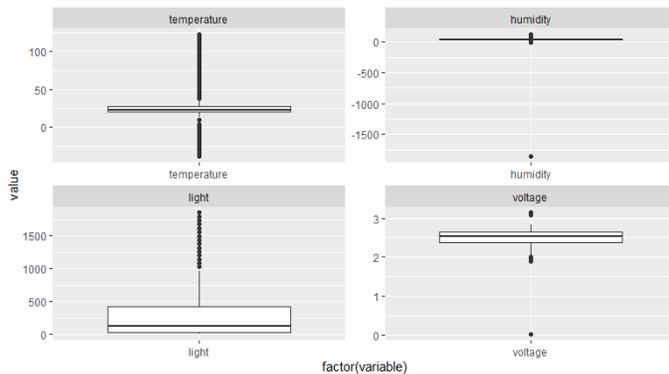


Fig. 3. Outliers.

The outliers in the master data are depicted in Fig. 3.

$$IQR = Q3 - Q1 \quad (3)$$

$$\text{Low outliers: } Q1 - 1.5IQR \quad (4)$$

$$\text{High outliers: } Q3 + 1.5IQR \quad (5)$$

Outliers in the data-set are replaced by their corresponding nearest quantile values. Data values outside the upper-boundary (high outliers) are replaced with corresponding third quantile values, similarly, data values outside lower-boundary (low outliers) are replaced with corresponding first quantile

values. Outlier treatment is carried-out for all key variables. Outlier detection is an essential step in data analysis since the un-treated outliers can affect the model results and predictions. Generally, outliers can be treated, suppressed or amplified. Our approach is to treat outliers as detailed above. Fig. 3 shows the outlined values for the four attributes which are in black colour.

The next step is weather prediction using multiple linear regression technique.

Multiple linear regression formula is:

$$y = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n + \epsilon \quad (6)$$

In equation 6, relying parameter predicted value is y , β_0 defines y -intersection (y -value with other variables made 0). A predicted ‘ y ’ value change with an increase in independent variable is given by $\beta_1 X_1$ or the first independent variable (X_1) with regression coefficient (β_1). This step of predicting y is repeated for all remaining independent variables to be tested. Finally the regression coefficient of the last independent variable is $\beta_n X_n$. Error present in the model is denoted by ϵ .

Important parameters required for identifying a best-fit line to each independent value in multiple linear regressions are:

- Coefficients resulting least error.
- t-statistic of the model. Z.
- The p-value.

t-statistic and p-value for each regression coefficient in the model is calculated and compared to determine statistical significance of the variable on outcome plus the magnitude of effect on outcome variable(y).

Regression analysis involves identification of residual data characteristics by means of assumption tests before Model build. Assumption tests are explained in the following subsection followed by model build. Regression analysis is essentially a parametric method and hence, validating assumptions is important. If underlying assumptions are violated Model results will not be accurate and predictions will be more prone to errors.

B. Regression Analysis Assumption Test

This section depicts assumption tests, which should be validated before regression analysis.

1) *Linearity*: Ideally, no fitted patterns are shown in the residual plot. It means, the red straight line shown in Fig. 4, should be approximately horizontal at zero. The model is found linear with the existence of a pattern, or a possible non-linear relationship. Here, there is no definitive pattern, and a linear relationship between the independent and dependent values seen, hence linear model suits.

2) *Normality of residuals*: For normality check, residuals can be visualized using QQ plot. As per normality assumption the residuals plot should be a straight line. In the given data, all the observations follow the defined reference straight line as shown in Fig. 5; hence normality assumption can be made.

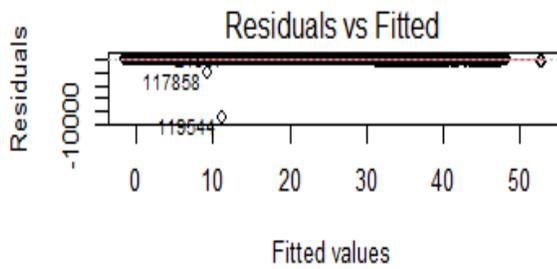


Fig. 4. Linearity.

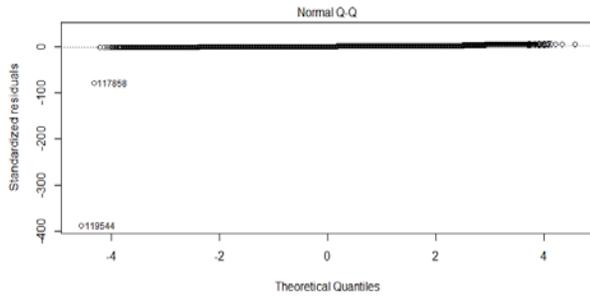


Fig. 5. QQ Plots.

3) *Homoscedasticity*: Fig. 6 shows how residuals are evenly spread along the range of predictors. The red line in the plot is nearly horizontal with similar dense distribution of points on either side. We can assume homogeneity of variance.

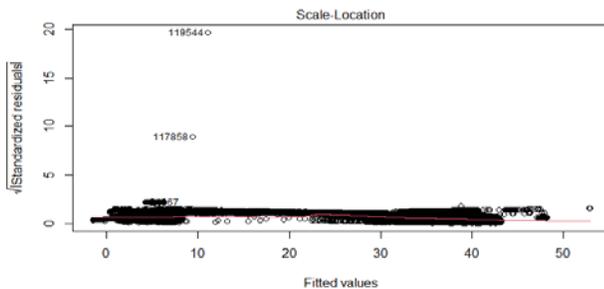


Fig. 6. Homoscedasticity.

4) *Durbin-Watson (DW) test*: DW test examines whether the error terms are autocorrelated. Null hypothesis states that no autocorrelation exists. The statistical DW test was performed and based on the p-value, we conclude that no autocorrelation exists. Statistical DW test yields the test result as ~ 1.9 , which means, no autocorrelation exists. Hence, this assumption is validated.

5) *VIF for multicollinearity*: On examining variable inflation factors for predictor variables, it was found that they do not exceed 5, hence, no multi-collinearity exists in the data set, which means, and the assumption of no multicollinearity is validated.

6) *Residual v/s Leverage*: The most influential 3 values are shown on the plot in Fig. 7; however, they can be exceptions, or, outliers. In this case, data does not present any high influence points.

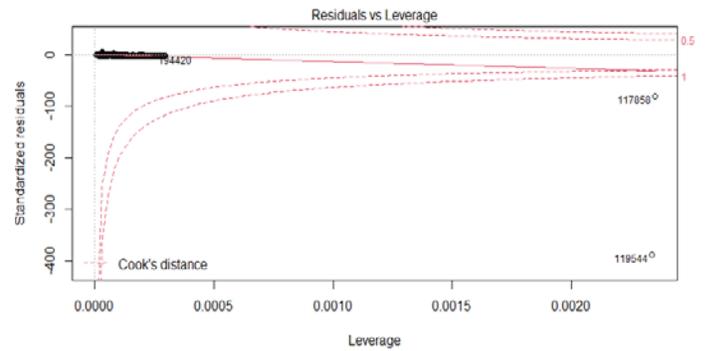


Fig. 7. Residuals vs Leverage.

7) *Cook's distance*: This score considers the combined values of leverage and residual parameters to determine an influential value. Regression analysis results will change with the inclusion or exclusion of influential value. An influential value has a larger residual. In linear regression analysis all outliers (or extreme data points) are not significant. Residuals show nearly even spread along the range of predictors. The red line in the plot is nearly horizontal with similar dense distribution of points on either side. We can assume homogeneity of variance.

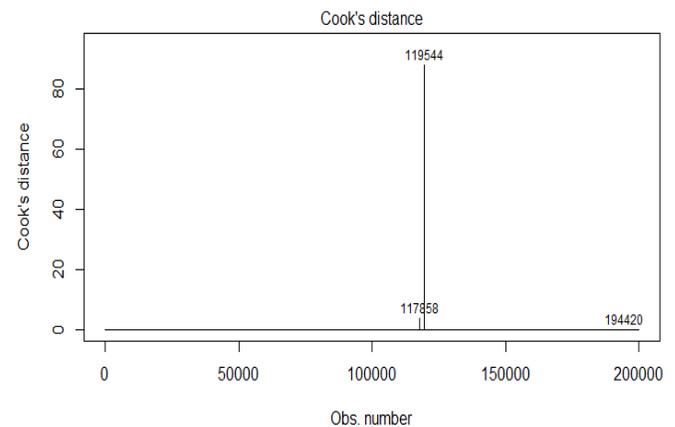


Fig. 8. Cooks Distance.

Cook's distance aids in determining the influential value. Here the thumb rule is, observations will have larger influence, if Cook's distance is more than $4/(n - p - 1)$ [14]. In the expression n represents the count of observations and p - the number of predictor variables.

In the given data, Cook's distance is too small as depicted in Fig. 8 and does not have significant influence on regression analysis. However, outliers, if any, must be detected and suitably handled.

C. Multiple Linear Regression Model

MLR is a statistical approach for predicting the output of a dependent variable by considering multiple independent variables [15]. MLR is capable of building a linear relationship between predictor variables and response values. Data is collected from the Intel Research Labs. Initially to eliminate noisy values, data pre-processing is done, avoiding reduction in prediction accuracy. Now, pre-processed data must be divided

into training and test data. The proposed algorithm needs to be trained utilizing training data for establishing relationship with several parameters. The final model will predict outcome of any new given data set containing data for same independent variables.

1) *Evaluation parameters:* The Model built is evaluated using various statistical metrics as listed below.

a) *R-square:* In MLR models r-square is used to measure a goodness-of-fit. Purpose of using this statistic is to judge how independent variables can mutually explain the dependent variable variance in percentage.

$$R - \text{Square} = \frac{\text{Variance explained by full model}}{\text{Total Variance}} \quad (7)$$

R-square increases every time a new independent variable is added to the Model. While a higher R-square is desirable, one has to vary about over-inflating the results and overfitting the Model.

b) *Adjusted R-square:* In regression Models, this statistic is used for comparing the goodness-of-fit with independent variables. The number of terms in the Model is adjusted with adjusted r-square. Importantly this parameter is mainly used to check an improvement in the Model fit with a new term. The adjusted R-squared value will automatically decrease whenever the new term fails to enhance the model fit by an adequate amount.

In our situation, r-square and adjusted-r-square are above-average values, and acceptable.

c) *Mean Absolute Error (MAE):* While predicting a set of values MAE is used to measure errors average magnitude, independent of direction.

With all equal weighted individual differences, MEA is computed as the test samples average parameter, considering the absolute differences between actual and prediction observations.

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (8)$$

Where, n – Number of samples, y_j -Expected value and \hat{y}_j -Predicted value.

d) *Root Mean Squared Error (RMSE):* RMSE is a squared output used for measuring average score of error.

It's the square root of the average of squared differences between prediction and actual observation.

$$RMSE = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (9)$$

Where, n – Number of samples, y_j -Expected value and \hat{y}_j -Predicted value.

MAE and RMSE values lie between 0 and ∞ , and are independent of the direction of errors. Both the metrics can be used for error prediction, expressing the variables in the interest of values units. For efficient modeling lower values are preferred, since both MAE and RMSE yield negative results.

Usually squared values of errors are taken before averaging, since RMSE gives relatively more weight to larger errors. Hence, RMSE is most suitable for large number of undesirable errors. Though RMSE penalizes larger errors, MAE has better interpretability, hence is considered.

The size of the dataset was decided by extracting random samples from the master dataset which keeps the distribution at a defined significance level. Basically, in order to achieve higher statistical significance, a machine learning model must be trained for larger dataset. But to save time sub-samples are selected maintaining other statistics same. Data distribution is fairly normal.

IV. RESULTS

All data pre-processing, exploratory analysis, Model build, tune and validation were performed using R language. In order to predict humidity, data pre- processing followed by multiple linear regression method was used. Existing data mining methods worked on homogeneous data, but the presented model is capable of handling heterogeneous data. The model performance was evaluated by using statistical metrics like R^2 , MAE, RMSE, etc.

1) *Intel dataset:* For experimentation, freely available Intel Lab dataset [16] was used. This dataset has nearly 2.3 million records collected by deploying 54 sensors in the Intel Berkeley Research lab. Sensors used were Mica2Dot, capable of collecting time-stamped weather information. The values are recorded by in-network query processing TinyDB system and these are recorded once every 31 seconds with humidity, temperature, light and voltage as key variables. The dataset format is given by: date, time, epoch, mote ID, temperature, humidity, light, and voltage. All the sensors are numbered with ids ranging from 1-54. Some sensor's values are missing or approximated. These measured variables are represented as, temperature in degrees Celsius, humidity ranging from 0-100% and its temperature corrected relative humidity. Light is recorded in Lux (1 Lux is equivalent to moonlight, 400 Lux corresponds to a bright office, and 100,000 Lux is equivalent to full sunlight). Voltage is in the range 2-3, measured in volts. Lithium ion cell batteries were used for providing constant voltage to sensors for their lifetime. It is observed that voltage variations are highly correlated with temperature.

2) *Model Results and Evaluation Metrics*

a) *R-square:* 0.692: Higher the R-square value, better it is. This statistical value shows variation between the dependent and independent variables. R-square is 0.692, which is considered a good-fit. Independent variables are able to explain a large amount of variance in dependent variable. p-value :<2.2e-16 => Model is statistically significant.

b) *Metrics:* As observed in Table I, average error is ~11%, hence we say Model is accurate upto 89%. MAE and RMSE are very small, hence our Model is very good; predictions generated from this Model will be very accurate.

c) *Residuals:* Table II gives residuals statistics for outliers treatment.

TABLE I. STATISTICS OF RESIDUALS

MAE	MSE	RMSE	MAPE
0.113799	0.03070434	0.17522655	Inf.

TABLE II. STATISTICS OF RESIDUALS FOR TREATING OUTLIERS

Min	1Q	Median	3Q	Max
-0.93682	-0.08049	0.00572	0.08174	0.97949

d) *Coefficients:* Table III lists the model coefficients for variables temperature, voltage, light and their correlations.

TABLE III. MODEL COEFFICIENTS

Variables	Estimate	Std. Error	t Value	Pr(> t)
Intercept	0.936820	0.002421	386.889	< 2e-16 ***
Temperature	-0.919190	0.003173	-289.665	< 2e-16 ***
Voltage	-0.134744	0.003269	-41.224	< 2e-16 ***
Light	-0.080437	0.007696	-41.224	< 2e-16 ***
temperature:voltage	1.493132	0.013469	110.853	< 2e-16 ***
temperature:light	0.140049	0.010163	13.780	< 2e-16 ***
voltage:light	-0.008874	0.010172	-0.872	0.383
temperature:voltage:light	-1.228874	0.038373	-32.024	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1732 on 139994 degrees of freedom.

Multiple R-squared: 0.6919, Adjusted R-squared: 0.6918.

F-statistic: 4.49e+04 on 7 and 139994 DF, p-value: < 2.2e-16.

3) *MLR Model with Interaction effects:* From the Model statistics, it can be seen that most co-efficient are statistically significant when seen individually, as well as with interaction effects. The interaction term co-efficient are statistically significant, suggesting an implicit interaction relationship between predictor variables (other than voltage and light). Model equation can be represented as below:

$$\text{humidity} = 0.936 - 0.92 * \text{temperature} - 0.13 * \text{Voltage} - 0.08 * \text{light} + 1.49 * (\text{temperature} * \text{voltage}) + 0.14 * (\text{temperature} * \text{light}) - 1.22 * (\text{temperature} * \text{voltage} * \text{light})$$

V. CONCLUSION

Changes in weather affect lives of various living beings. Main idea of this paper is to analyze the data collected from WSNs of Intel Lab and make appropriate decision in order to convey right information at right time to help save lives.

Machine Learning techniques are used for weather prediction, and MLR algorithm is built using temperature, humidity, light and voltage as the key variables. We have evaluated the Model and model results are documented above. The resultant model can predict with high degree of accuracy and can be expanded for further work. Values of statistical parameters indicate that the proposed model is statistically more significant compared to other existing techniques.

REFERENCES

- [1] Mohd Fauzi Othmana, Khairunnisa Shazalib, 2012, 'Wireless Sensor Network Applications: A Study in Environment Monitoring System', International Symposium on Robotics and Intelligent Sensors 2012 (IRIS 2012), Procedia Engineering 41 (2012) 1204 – 1210.
- [2] Munmun B, Tanni D, Sayantanu B (2018), 'Weather forecast prediction: an integrated approach for analyzing and measuring weather data'. Int J Computer Appl. <https://doi.org/10.5120/ijca2018918265>.
- [3] Kothapalli S, Totad SG (2017), 'A real-time weather forecasting and analysis'. In: IEEE international conference on power, control, signals and instrumentation engineering (ICPCSI-2017), pp 1567–1570.
- [4] Almgren K, Alshahrani S, Lee J (2019), 'Weather data analysis using hadoop to mitigate event planning disasters'. <https://scholarworks.bridgeport.edu/xmlui/handle/123456789/1105> Accessed 1 Feb 2019.
- [5] N. Krishnaveni, A. Padma, 2020, 'Weather forecast prediction and analysis using sprint algorithm', © Springer-Verlag GmbH Germany, part of Springer Nature 2020, Journal of Ambient Intelligence and Humanized Computing <https://doi.org/10.1007/s12652-020-01928-w>.
- [6] Taksande AA, Mohod PS (2015), 'Applications of data mining in weather forecasting using frequent pattern growth algorithm'. Int J Sci Res (IJSR) 4(6):3048–3051.
- [7] Wang Z, Mujib ABM (2017), 'The weather forecast using data mining research based on cloud computing'. J Phys 1:1–6.
- [8] Salman AG, Kanigoro B, Heryadi Y (2015), 'Weather forecasting using deep learning techniques'. In: 2015 International conference on advanced computer science and information systems (ICACSIS), pp 281-285.
- [9] Maqsood I, Khan MR, Abraham A (2004), 'An ensemble of neural networks for weather forecasting'. Neural Comput Appl 13(2):112–122.
- [10] Oury DTM, Singh A (2018), 'Data analysis of weather data using hadoop technology'. In: Smart computing and informatics, Springer, Singapore, pp 723–730.
- [11] Manogaran G, Lopez D (2018), 'Spatial cumulative sum algorithm with big data analytics for climate change detection'. ComputElectrEng 65:207–221.
- [12] Mahmood MR, Patra RK, Raja R, Sinha GR (2019), 'A novel approach for weather prediction using forecasting analysis and data mining techniques'. In: Saini H, Singh R, Kumar G, Rather G, Santhi K (eds) Innovations in electronics and communication engineering. Lecture notes in networks and systems, vol 65. Springer, Singapore.
- [13] Madden, S. (2004). Wireless sensor data [Intel lab data]. Retrieved from <http://db.lcs.mit.edu/labdata/labdata.html>.
- [14] Peter Bruce and Andrew Bruce, Practical Statistics for Data Scientists, Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472, Copyright © 2017.
- [15] C. G. N. De Carvalho, D. G. Gomes, J. N. De Souza, and N. Agoulmine, "Multiple linear regression to improve prediction accuracy in WSN data reduction," in Network Operations and Management Symposium (LANOMS), 2011.
- [16] Intel Lab Data. <http://berkeley.intel-research.net/labdata/>, 2004.

Detecting Diabetic Retinopathy in Fundus Images using Combined Enhanced Green and Value Planes (CEGVP) with k-NN

Minal Hardas, Sumit Mathur, Anand Bhaskar
Electronics and Communication Engineering
Sir Padampat Singhanian University (SPSU), Udaipur, India

Abstract—Diabetic Retinopathy (DR) is a disease that causes damage to the blood vessels of the retina, especially in patients having high uncontrolled blood sugar levels, which may lead to complications in the eyes or loss of vision. Thus, early detection of DR is essential to avoid complete blindness. The automatic screenings through computational techniques would eventually help in diagnosing the disease more accurately. The traditional DR detection techniques identify the abnormalities such as microaneurysms, hemorrhages, hard exudates, and soft exudates from the diabetic retinopathy images individually. When these abnormalities occur in combination, it becomes difficult to predict them and the individual detection (traditional 4 class classification) accuracy decreases. Hence, there is a need to have separate combinational classes (16 class classification) that help to classify these abnormalities in a group or one by one. The objective of our work is to develop an automated DR prediction scheme that classifies the abnormalities either individually or in combination in retinal fundus images. The proposed system uses Combined Enhanced Green and Value Planes (CEGVP) for processing the fundus images, Principal Component Analysis (PCA) for feature extraction, and k-nearest neighbor (k-NN) for classification of DR. The suggested technique yields an average accuracy of 97.11 percent using a k-NN classifier. This is the first time that a 16-class classification is initiated that precisely gives the ability and flexibility to map the combinational complexity in a single step. The proposed method can assist ophthalmologists in efficiently detecting the abnormalities and starting the diagnosis on time.

Keywords—Combined enhanced green and value plane; diabetic retinopathy; fundus image; image processing; k-NN; principal component analysis

I. INTRODUCTION

Diabetes is an epidemic affecting millions of people worldwide [1]. DR is a chronic retinal disorder that is caused by the long-term impact of diabetes mellitus [2]. It is a disease that is characterized by gradual progressive alterations in the retinal microvasculature. People with diabetes and less controlled blood sugar are likely to suffer from DR. It occurs when high blood sugar levels damage the walls of small blood vessels in the retina [3]. These vessels can swell, leak, or close, stopping blood from passing through them. Frequently, there is an accumulation of fluid in the part of the retina called macular edema [4]. Sometimes, in more advanced cases, the supply of blood to the retina is cut off, which results in the growth of abnormal new blood vessels called neo-

vascularization. These new fragile vessels can bleed, creating vision-impairing hemorrhages, scar tissue, and separation of the retina from the back of the eye. In DR, blood vessels leak fluid and blood on the retina. These vessels form features such as microaneurysms, hemorrhages, hard exudates, and soft exudates or cotton-wool spots [5]. The microaneurysms are hypercellular saccular out pouching of the capillary wall. They appear as deep-red dots varying from 25 to 100 μm in diameter and have distinct margins. Retinal microaneurysms are usually the first ophthalmoscopic sign of DR [6]. They are located predominantly within the inner nuclear layer and in the deep retinal capillary network. Microaneurysms are the hallmark of Non-Proliferative Diabetic Retinopathy (NPDR) [7]. Intraretinal hemorrhages are another predominant feature of NPDR. It results from ruptured microaneurysms, capillaries, and venules, and is mostly within the outer plexiform and inner nuclear layers. Retinal hemorrhages are blot-shaped or flamed-shaped. Hard exudates are an ophthalmoscopic feature of background diabetic retinopathy [8]. They result from an increase in vascular permeability and the leakage of fluid and lipoprotein in the surrounding tissue. The hard exudates are fat-filled (lipoidal) histiocytes. They are small white or yellowish-white deposits with sharp margins in the outer layers of the retina, deep in the retinal vessels. Soft exudates or cotton-wool spots are localized infarctions of the nerve fiber layer with secondary coagulative necrosis of the retina [9]. They appear as pale yellow or white lesions with ill-defined edges in the superficial retina. The presence of soft exudates is the symbol for the onset of progressive change in DR. A cotton-wool spot can occur singly or in conjunction with hemorrhages and microaneurysms and represent retinal microvasculopathy.

The retinal appearance of diabetes mellitus is broadly classified as Non-Proliferative Diabetic Retinopathy (NPDR) or Proliferative Diabetic Retinopathy (PDR). NPDR occurs when there are only intraretinal microvascular changes, such as altered retinal vascular permeability and eventual retinal vessel closure [10]. In advanced NPDR, non-perfusion of the retina may develop and lead to the proliferative phase. The PDR is characterized by the formation of new vessels.

Contrast Limited Adaptive Histogram Equalization (CLAHE) is the process that reduces the noise amplification and it is the type of adaptive histogram equalization [11]. It operates on small regions of the image and works with neighboring tiles to avoid noise enhancement. Top hat filter is

a morphological filtering operation that works on grayscale images [12]. This filter relies on the structuring element size and, when fine tuned, it can increase the visibility of the features like red-colored veins in a fundus image. Tunable top hat filters are used for DR detection and when combined with CEGVP, it improves the accuracy of the proposed system.

Although some work has been done in the past, none of the methods have shown combinational identification of the abnormalities that are present in the retinal fundus image. It is quite desirable to have a combinational class because individual detection of these abnormalities may lead to false detection during segmentation, affecting the performance of the system. Hence, a separate class was required for predicting the combined abnormalities, as most of them occur in combination. The major contribution of the proposed method is 16 class classification using CEGVP, which is further passed through machine learning algorithms to diagnose DR. The CEGVP output was further improved using CLAHE and top hat and bottom hat approach. The features were then extracted using PCA and a 16 class classification of the abnormalities was carried out using a k-NN classifier.

The following sections of the paper are organized as follows: Section II presents an explicit literature review, Section III describes the methodology, Section IV highlights the results and discussions. Finally, Section V concludes the study.

II. LITERATURE REVIEW

The researchers have investigated and developed several algorithms for detection of DR. A new technique was developed by Lachure et al. [13] for diagnosing PDR and NPDR. Their work consisted of detecting the abnormalities such as microaneurysms using morphological opening operations and exudates using morphological closing operations. The splat and Gray Level Co-occurrence Matrix (GLCM) features such as entropy, contrast, homogeneity, and energy were extracted and the image was further classified into PDR and NPDR using machine learning classifiers. The performance of SVM was better than compared to k-NN classifier. Safitri et al. [14] have proposed a new method for classifying the DR into three grades. Their work comprised segmentation of green channel image, applying morphological and masking operations, and computing the values of fractal dimensions using the box-counting method. These values were analyzed and classification of DR was performed using the k-NN classifier. The proposed method provided the best accuracy of 89.17% for $K=3$ and $K=4$. Labhade and his team [15] had presented an automated method for the detection of DR. Their work included preprocessing, feature extraction, and classification using machine learning algorithms. The preprocessing comprised grayscale conversion and histogram equalization for enhancing the contrast of grayscale fundus images. Textural features were extracted using GLCM by considering 3 angles of 0, 45, 90 degree, and 2 distances. The statistical moments were computed and the retinopathy grade between 0-3 was assigned for each of the fundus images. The SVM classifier provided better accuracy of 88% as compared with other classifiers. Kushol and his group [16] had presented a new blood vessel enhancement technique that separates the

blood vessels from the background image. The top-hat and bottom-hat transformations with optimal structuring elements were used to enhance the image. The proposed method yielded an average accuracy of 0.9379 and 0.9504 on DRIVE and STARE datasets, respectively. Kaur et al. [17] had proposed a reliable method of exudate segmentation using dynamic decision thresholding in the diagnosis of DR. Their work includes enhancement of retinal image, segmentation, and elimination of anatomical structures such as optic disc and blood vessels, and segmentation of exudates using adaptive image quantization and dynamic decision thresholding process. The proposed method resulted in a mean sensitivity, specificity and accuracy of 94.62%, 98.64%, 96.74%, respectively, at image based evaluation. Marin and the team [18] had presented a feature-based supervised classification technique for detecting Diabetic Macular Edema (DME) in fundus images. Their work comprised detecting the exudates by using digital image processing algorithms. Edge strength-based features and the features based on responses from the Gaussian and Difference of Gaussian (DoG) filter bank were computed. The final detection of exudate was obtained by considering only the regions whose probability exceeded a certain threshold value and the severity was graded using supervised classification techniques. The k-NN and SVM classifier resulted in an accuracy of 0.955 and 0.97422 for diagnosing the retinopathy disease. Issac and his team [19] have developed a method for detecting red and bright pathologies for diagnosing DR. A normalization process followed by anisotropic diffusion was used for segmenting bright lesions. A shade-corrected green channel image along with morphological flood filling and regional minima operations were used for detecting the red lesions. A quantitative analysis was performed to grade the severity of the disease. The proposed method using SVM based classifier obtained an average accuracy of 92.13% with a sensitivity of 92.85% and specificity of 80% on the DIARETDB1 dataset. A novel method was presented by Chetoui and his group [20] that comprised texture features, namely Local Ternary Pattern (LTP) and Local Energy-based Shape Histogram (LESH) for diagnosing DR. The histogram was computed using the extracted features and the performance of the system was evaluated for various kernels of SVM classifier. It was observed that LESH outperformed best in terms of accuracy and ROC characteristics. The accuracy of 0.904 and ROC of 0.931 was obtained using the SVM Radial Basis Function kernel. Amin and his team [21] had developed an automated method for the detection and classification of DR using hybrid features. Their work included lesion enhancement using the local contrast method, optic disc elimination, and lesion segmentation. The geometrical and statistical features were extracted for each candidate lesion. Abnormal and normal images were differentiated using multiple classifiers. The proposed method validated on the DIARETDB1 dataset resulted in an accuracy of 92.3% using k-NN and SVM classifier. Sahu et al. [22] have proposed denoising of fundus images using CLAHE. They could achieve an improvement of 7.85%, 1.19%, 0.12%, and 1.28%, in Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Correlation coefficient (CoC), and Edge preservation index (EPI), respectively, over the existing methods. Huda et al. [23]

proposed an automated approach for diagnosing DR. Their study contained normalization, data standardization, and feature extraction from the retinal samples. The feature importance property was used to select the important features from each sample. The scores of each feature were generated and a higher score feature was selected. Their method resulted in six class classifications of the disease and their performance was evaluated using different classifiers. The proposed method with a logistic regression model achieved a precision of 97% and recall of 92%. Reddy et al. [24] have introduced an ensemble-based machine learning model for detecting DR. The grid search parameter tuning optimization method was used for choosing the optimal hyper parameters and the voting mechanism was used to make predictions for each classifier model. It was concluded that the ensemble machine learning model outperforms the individual machine learning algorithms. A study conducted by Alabdulwahhab and his team [25] had introduced a novel method for detecting DR. Their work comprised identifying the most discriminative interpretable features using socio-demographic and clinical information. The factors such as HbA1c, duration of diabetes, body mass index, systolic blood pressure, and the use of medication were used to discriminate the DR patients. Thus, a combination of ophthalmology and ML was integrated for diagnosing the disease pattern. The random forest classifier outperformed best by accurately classifying 86% of the DR patients. The author Sharma and his team [26] have developed an automated system for detecting DR using a combination of image processing and machine learning. Their work included pre-processing techniques such as gray scale conversion, canny edge detection and morphological operations to obtain a clear fundus image. The statistical features were extracted from the images and the performance of the system was evaluated using various classifiers. The weighted k-NN provided an accuracy of 85.8%, SVM with 87.2% and decision tree with 88.6%.

The literature review has highlighted the implementation of several machine learning algorithms applied independently on retinopathy datasets for detection of DR. As reviewed, there was a maximum of 6 class classification performed for diagnosing DR. Each of these techniques contributed to an individual detection of the abnormalities in the fundus images because of which the performance of the system was affected. The authors could achieve a maximum accuracy of 95.55%. The sensitivity computed for most of the methods was between 90% to 93%. AUC of 1 was obtained for very few methods. None of the techniques could identify the combination of the abnormalities as most of the time these abnormalities occur in combination in the retinal fundus images. Hence, a 16 class classification comprising a unique combination of CEGVP and dimensionality reduction approach using PCA is proposed that accurately classifies the four abnormalities either individually or in a group for diagnosing DR.

III. METHODOLOGY

The proposed method consists of classifying four types of abnormalities, such as hard exudates, soft exudates,

microaneurysms and hemorrhages, either individually or in combination from a fundus image using CEGVP and k-NN classifier. Fig. 1 shows the conceptual diagram of the proposed system. A fundus camera captures the images of an eye, which is then processed using CEGVP. The features are extracted and fed into the k-NN classification algorithm, which classifies each image into one of the 16 categories. All 16 categories represent various abnormalities that may or may not be present in a fundus image. The diagnosis can be made based on the category into which the k-NN algorithm classifies the given test image.

The suggested technique combines four steps: preprocessing, segmentation, feature extraction, and classification. The input image is converted to grayscale. This grayscale image is used for a mass generation and binarization. The green color plane is enhanced by stretching the histogram and applying the mask. CLAHE is used to extract blood vessels from images using the Lab color space and exudates are extracted using top hat filtering. After enhancement, the combined green and value plane is passed through a Canny edge detector and subjected to morphological operations for object detection after binarization. Finally, the binary objects are combined, and the principal component is extracted as a feature for the k-NN algorithm. Fig. 2 shows the overall process of the proposed work.

The first step of processing comprises reading the original colored image (Fig. 3(a)) from the database and converting it into grayscale (Fig. 3(b)). The height and width of the input image was compared and the larger dimension was set to 640 and the other dimension was scaled, such that the aspect ratio of the image was maintained (Fig. 3(c)). A binary image (Fig. 3(d)) was generated from the grayscale image by using a threshold of 0.05 on the scale between 0 and 1. The resized original input image (colored) was again converted into a grayscale image. Each pixel of the new grayscale image was checked. If the value of a pixel was less than 10, then that pixel value was set to 0. Otherwise, if the value of a pixel was greater than 10, then that pixel value was set to 255, achieving the mask for the fundus (similar to Fig. 3(d) but with the dimension of 640 pixels on the larger dimension).

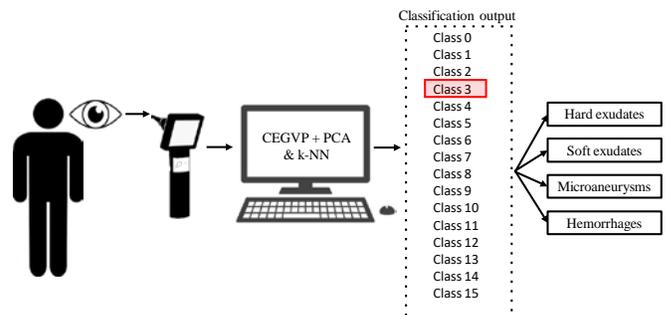


Fig. 1. Conceptual Diagram of the Proposed System. The Eye Images Captured by the Fundus Camera are Processed using CEGVP. The PCA Features are Extracted and then Fed to k-NN Classifier. Each Class out of 16 Classes Represents a Unique Combination of Abnormalities that may be Present in a Fundus Image.

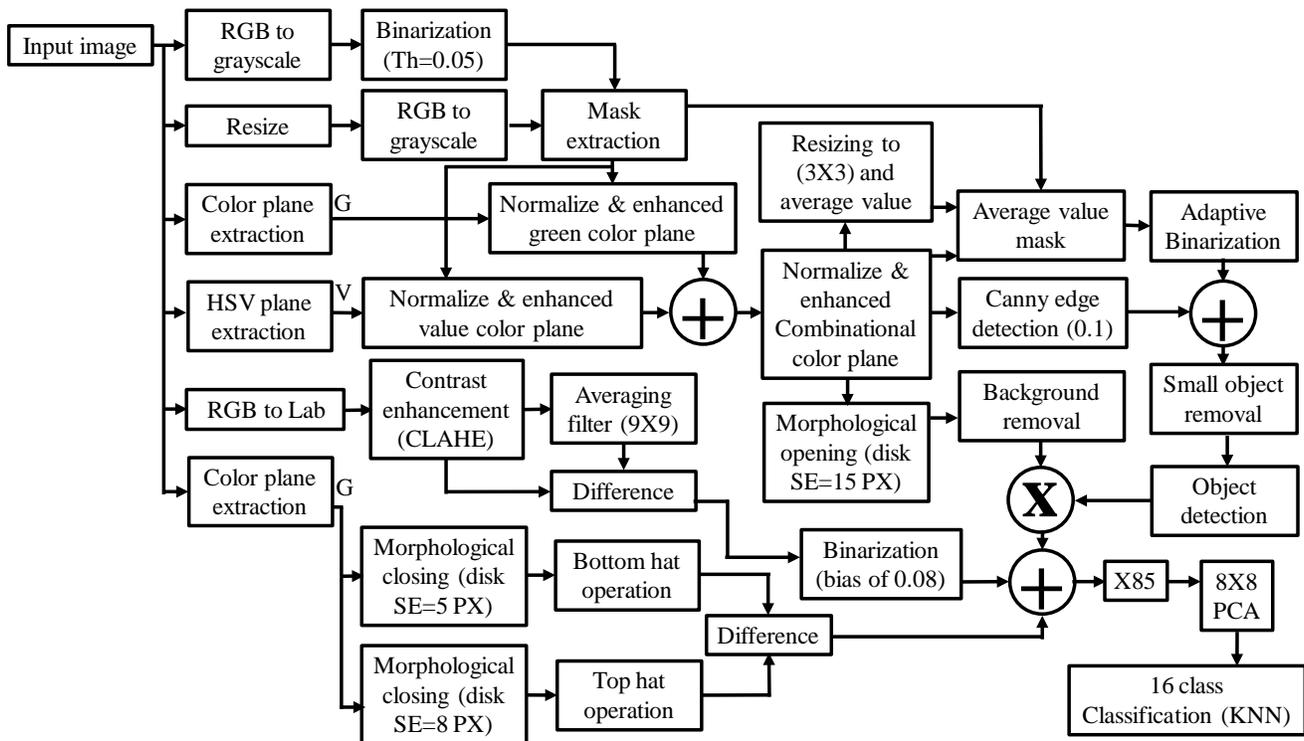


Fig. 2. Proposed Methodology for Extraction of different Abnormalities from a Fundus Image using CEGVP and k-NN. The Input Image is converted to Grayscale as well as HSV, Lab, and RGB Color Planes are extracted. The Green Color Plane is enhanced using Histogram Stretching and Application of the Mask. Lab Color Space is used to extract the Blood Vessels from an Image. Top Hat Filtering is used to Extract Exudates. The CEGVP is Passed through a Canny Edge Detector and Subjected to Morphological Operations Post Binarization for Object Detection. Finally, all the Binary Objects are merged and the Principal Component is extracted as a Feature for the k-NN.

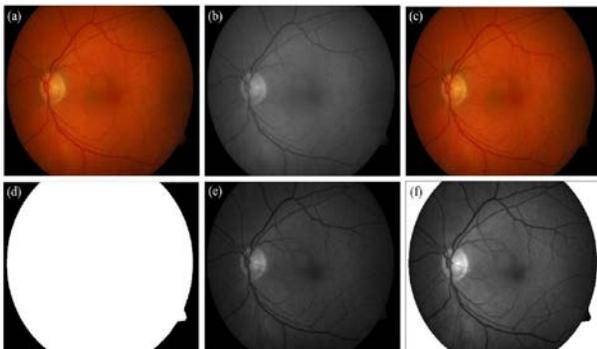


Fig. 3. Output at Various Stages of CEGVP for the Fundus Image. (a) Original Input (b) Grayscale Converted (c) Resized (d) Background Masked Image with a Threshold of 0.05 (e) Extracted Green Color Plane (f) Enhanced and Masked Green Color Plane with Background Removal.

Red and green color planes were extracted from the original input image with green color plane as shown in (Fig. 3(e)). A structuring element was defined that had a size of 350 pixels and a shape of a disk. Morphological opening operation was performed on the green color plane. For the morphological operation of opening previously defined structural element was used on the green color plane. The output of the opening operation was then subtracted from the original green color plane. A maximum pixel value of this subtracted image was found out. The mask image was then used for masking the outer region with white color (pixel value of 255) as seen in (Fig. 3(f)). A copy of this processed image was made for further processing. All the dark pixels in

the green channel were preserved whereas all the lighter pixels with a value greater than 50 were whitewashed. The copy image of the green channel was then contrast stretched between 0 and 255 (Fig. 3(f)).

The original image was used yet another time to extract the red, green, and blue color planes. A mask was applied to the green color plane and the same color plane was contrast enhanced for maximum value (255). A secondary mask was created with the help of the green channel having the threshold of 60 on a scale of 0 to 255. All the pixel values below 60 were made 0, and 255 otherwise. A blue color mask was used, similar to green color with a threshold of 130. A similar operation was performed on the blue color plane. The color pane of the original image was changed to HSV from RGB and further hue, saturation, and value planes were separated. The value plane was then masked, and contrast-enhanced between 0 to 255 (Fig. 4(a)). Similar operations were performed on saturation and hue color planes. The hue color plane was enhanced 18 times. The combinational color plane was generated using the enhanced green plane and the value color plane (Fig. 4(b)). Further, all the max values were replaced with mean values to ensure that contrast enhancement can be done properly. Because of this operation, all the regions of the mask were neglected. The combinational plane was then contrast enhanced (Fig. 4(c)) and the entire image was rescaled to a smaller matrix of 3X3 to extract a single-valued function available at the center of these 3X3 images (Fig. 4(d)).

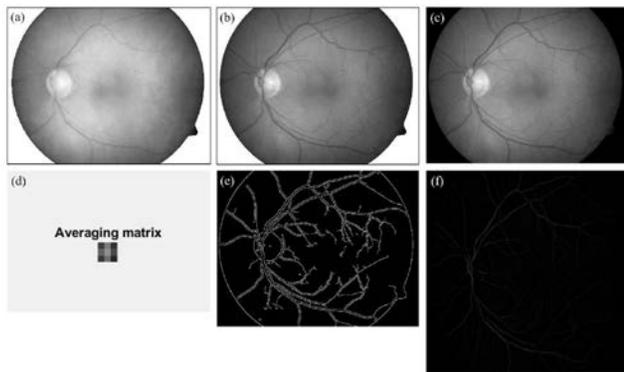


Fig. 4. Outputs at different Stages for the Input Image under Consideration. (a) Masked and Normalized Value Color Plane. (b) Combinational Green and Value Color Plane. (c) Enhanced Combinational Green and Value Color Plane. (d) Averaging Matrix with 3 X 3 Dimensions. (e) Canny Edge Detector. (f) Morphological Bottom Hat Operation.

The combination image of the green color plane and value color plane was masked using the mask image and further enhanced with the saturation values at the bottom and top 1% amongst all pixel values. The image was rescaled to 3X3 and the entire process was repeated. Finally, a Canny edge detector was applied to this enhanced combinational color plane (green & value plane) with a factor of 0.1 (Fig. 4(e)). Simultaneously, adaptive binarization was applied to the enhanced combinational color plane (green & value plane). All pixels in the output of the binarization process that had a value of less than 200 were removed. A structuring element of disk shape with size 15 pixels was used to call a morphological opening operation on the enhanced combinational color plane. Further, the background was then removed using the image which was operated using the opening operation. The overall contrast was then enhanced and all the objects touching the border were removed. A 4-pixel neighboring connectivity was checked for connected components and areas of all the objects segmented from this neighboring pixel connectivity were used to find the maximum area. All the segregated objects were then binarized and the entire image was resized to 512 X 512 pixels. This resizing ensured that machine-to-machine variability is not affecting the classification results. The original input image was then passed through the color channel separator to segregate the green color channel component. A structuring element with a disk shape of 5 pixels was then used on the green channel to achieve bottom hat filtering (Fig. 4(f)).

The same green channel image was also simultaneously passed through a top-hat filter (Fig. 5(a)) but with a structuring element of disk shape having a size of 8 pixels. Finally, a bottom hat filtered image was subtracted from the top-hat filtered image, and binarization with a threshold of 0.1 was carried to extract the hard exudates (Fig. 5(b)). The original image was resized to 584 X 656 pixels and converted into a Lab color space to extract the principal components from the converted image using PCA. The entire image was then normalized and contrast enhanced using CLAHE (Fig. 5(c)). An averaging filter of 9 X 9 pixels was used to exclude the background. Thresholding was performed on the background removed image and then the entire image was converted to binary with a level of 0.08% lower than the

automated thresholding level to extract the blood vessels as shown in (Fig. 5(d)). Small objects less than 100 pixels were removed and then the image was inverted. Finally, the entire image is resized to 512X512 to merge with the earlier results. All the three outputs detecting microaneurysm, hemorrhages, and exudates in their binary form were added and scaled by a scalar of 85 to produce a final grayscale output (Fig. 5(e)) which can then be given to the feature extraction algorithm. The final image was then passed to extract the first 64 components. Finally, a class is assigned to these 64 extracted components based on the supervisory dataset DIARETDB0 [27], DIARETDB1 [28], and Indian Diabetic Retinopathy Image Dataset (IDRid) [29].

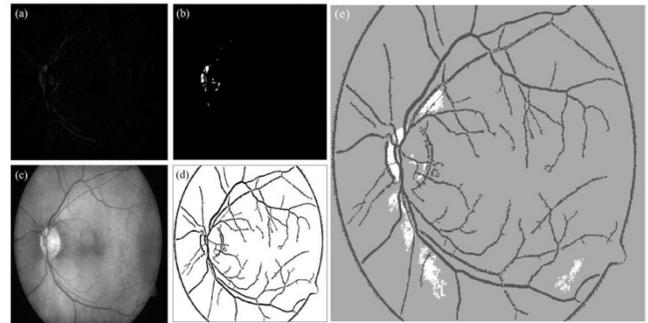


Fig. 5. Outputs at different Stages of CEGVP. (a) Morphological Top-Hat Operation (b) Hard Exudates (c) Contrast-enhanced Lab Color Space using CLAHE (d) Blood Vessel Detection (e) Combined Output from CEGVP, Morphological Top-hat, CLAHE, and Canny Edge Detector. The Image is converted to Grayscale with Four different Levels: 0, 85, 170, and 255.

The novelty of the proposed methods lies in the combination of the green sub-color plane from RGB planes and the value sub-color plane from the HSV color planes (CEGVP) to enhance the overall features. This is the first time a unique combination of combined color planes and 16 class classification has been performed for predicting the abnormalities in the fundus image for detection of DR. This work presents the combination of image processing techniques for enhanced feature extraction and classical machine learning approach, such as k-NN for classification. The performance of the proposed method was tested using three publicly available databases. In all 300 images with DIARETDB0 (130 images), DIARETDB1 (89 images), and IDRid (81 images) were used respectively.

IV. RESULT AND DISCUSSION

A 5-fold cross validation was used for the analysis and experimentation purposes. The ratio in which the images were split was chosen to be 70% for training and 30% for testing. To verify the performance of the designed system, 450 images were chosen randomly from the test dataset.

A. Result Analysis

1) *Performance measures*: The performance measures such as accuracy, sensitivity, specificity and area under the curve (AUC) were used to evaluate the performance of the proposed method. Table I and Fig. 6 shows the performance metrics of various classifiers for PCA feature vector 8x8 using 16 class classification.

TABLE I. PERFORMANCE METRICS USING PCA FEATURE VECTOR 8x8

Classifiers	Metrics			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
KNN	97.11	98.47	91.64	1
SVM	95.77	93.66	97.43	1
Decision Tree	87.3	61.3	64.7	1

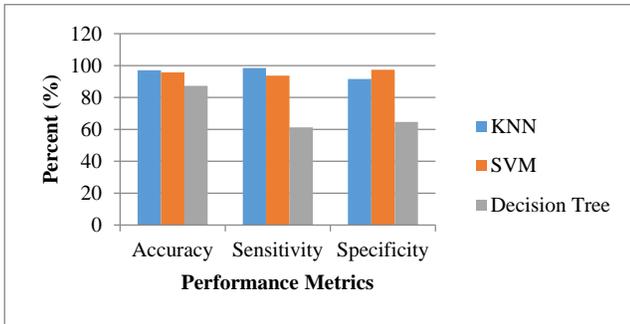


Fig. 6. Comparison of Performance Metrics using 8x8 PCA. k-NN Performed Best in Terms of Accuracy and Sensitivity as Compared to other Classifiers, Whereas in Terms of Specificity SVM was Most Superior amongst All.

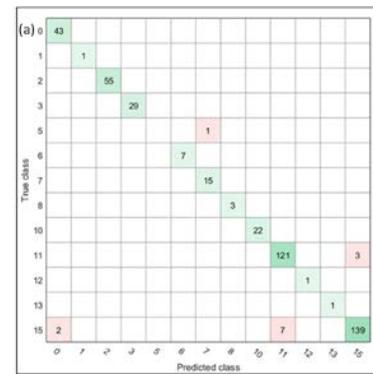
It was observed from the graph that the highest accuracy of 97.11% and sensitivity of 98.47% was obtained for k-NN classifier. SVM achieved the best specificity of 97.43% as compared to other classifiers. The decision tree classifier performed very low in terms of all the performance measures. The AUC for all classifiers was found to be 1. Thus, it was seen that k-NN was the most optimal classifier for the proposed system.

To increase the ability of the system, bag-of-words concept was used that provided more images for training and testing purpose. Fig. 7(a) shows the confusion matrix for 16 class classification using k-NN. The three classes 4, 9, and 14 were not found in the confusion matrix, as there were no images for training and testing for these classes in the datasets. The worst accuracy was found for class 5, where all the images were mapped to class 7. Other than classes 11, and 15, all the classifications were 100% accurate. Whereas for class 15 and 11, the number of images were more, which resulted in a fraction of misclassification, which is roughly 4.6%. There were maximum of 148 images of class 15, followed by 124 images of class 11 and 55 images of class 2.

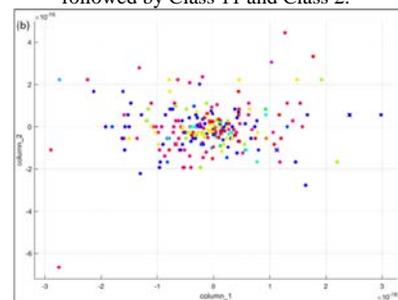
Fig. 7(b) shows the scatter plot after PCA features were extracted for the 16 classes and the top 2 features were plotted. It can be visually seen that no straight line can separate various classes from each other, and hence k-NN was the preferred method over the traditional SVM.

The Receiver Operating Characteristic (ROC) curve with the false positive and true positive rates is as shown in Fig. 7(c).

2) *PCA feature vector of optimal sizes vs accuracy*: To select the most optimal method of classification and the optimum size of the PCA feature vector, a separate study was carried out as shown in Table II and its output was summarized in the graph shown in Fig. 8.



(a). Confusion Matrix for 16 Class Classification using k-NN. The Images in the Categories 4, 9, and 14 were missing as there were no Images for these Particular Classes in the Datasets. Maximum Images were of Class 15 followed by Class 11 and Class 2.



(b). Scatter Plot for 16 Class Classification using k-NN. As Observed, it is Very Difficult to Segregate a Linear Line among the Various Classes and Hence k-NN is the most Suited Strategy.

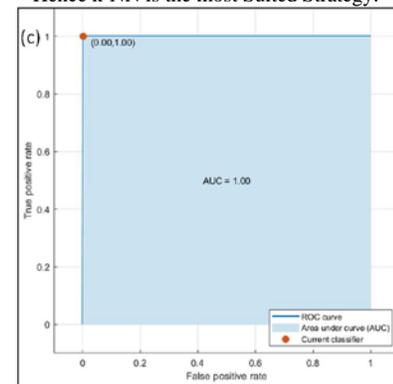


Fig. 7. (c). Area Under Curve (AUC) for 16 Class Classification using k-NN Class 15 has an Area under the Curve (AUC) of 1.0.

TABLE II. COMPARISON OF PCA FEATURE VECTOR WITH ACCURACY

PCA Feature Vector	Accuracy (%)		
	KNN	SVM	Decision Tree
1x1	91.6	39.6	57.3
2x2	96.4	90.2	68.7
3x3	96.9	96.9	81.3
4x4	96.7	96.4	83.6
5x5	95	95.8	86
6x6	95.6	95.3	85.9
7x7	96.9	96	86.9
8x8	97.1	95.8	87.3
9x9	95.3	95.2	86.6
10x10	94.1	94.1	86.2

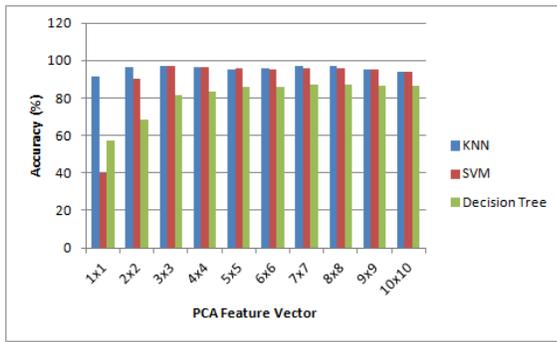


Fig. 8. PCA Feature Vector Size with Accuracy for SVM, k-NN, and Decision Tree Algorithms. The 8x8 PCA Vector Size was found out to be the Best for k-NN that had Maximum Accuracy. SVM Achieved Higher Accuracy only when PCA Vector Size was 3X3 but showed Poor Performance for PCA Feature Vector Size 1X1. The Decision Tree Always had approximately 10% Less Accuracy than k-NN for all the Feature Vector Sizes.

It was observed that all the algorithms do get affected if the feature vector size was varied. Since k-NN had the least variation, it was considered being less dependent on the size of the feature vector. The decision tree was highly dependent on the feature vector size. It has a bimodal distribution of accuracy making it difficult to recognize the correct PCA feature vector size.

The 8X8 PCA feature vector size was discovered to be the optimum vector size for k-NN and 3x3 for SVM, as highest

accuracy was obtained by both the classifiers for these vector sizes. SVM performed badly with less than 40% accuracy at PCA feature vector size 1 X 1. The decision tree, for all conceivable feature vector sizes always had approximately 10% less accuracy as compared to k-NN. As a result, 8X8 PCA feature vector was found to be the most suitable vector size for KNN classifier.

B. Discussions

Table III shows the comparative analysis of the proposed work with the existing approaches, as reviewed in the literature. The proposed method achieved the highest accuracy compared to all the similar methods that use k-NN or SVM classifiers. Marin et al. [18] have achieved a similar accuracy, but their sensitivity and specificity were poor. The sensitivity obtained using k-NN classifier was highest using the proposed technique. Lachure et al. [13] could achieve 100% specificity, but they have considered only three classes and overall accuracy was 90%. AUC for the proposed method was comparable to the others as reported in the literature, whereas the number of classes that are segregated in our work is at least 2.5 times than every method ever reported. The overall performance of k-NN is better as compared to SVM and hence we proposed KNN classifier for abnormality detection in fundus images. Some other methods perform better in terms of accuracy, but they are neither working on 16 class classification nor using KNN and SVM classifiers hence, they were excluded from the comparison table.

TABLE III. COMPARISON OF PROPOSED SYSTEM WITH EXISTING METHODS

Ref	Technique	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC	No. of classes	Classifier
Lachure et al. [13]	Abnormality detection using morphological operations, splat and GLCM features extraction.	90	90	100	...	3	SVM
Safitri et al. [14]	Segmentation by applying morphological, masking operations and computing the values of fractal dimensions using box-counting method.	89.1	4	KNN
Labhade et al. [15]	Textural features extraction using GLCM.	88.71	4	SVM
Kaur et al. [17]	Segmentation of exudates using adaptive image quantization and dynamic decision thresholding process.	87	91	94	...	2	SVM
Marin et al. [18]	Edge strength-based features and the features based on responses from Gaussian and difference of Gaussian (DoG) filter bank were computed.	95.55	90	70	...	2	KNN
Issac et al. [19]	Anisotropic diffusion for segmenting bright lesions and a shade-corrected image along with morphological operations for detecting the red lesions.	92.13	92.85	80	...	4	SVM
Chetoi et al. [20]	Texture features namely Local Ternary Pattern (LTP) and Local Energy-based Shape Histogram (LESH).	90.4	0.931	2	SVM
Amin et al. [21]	Lesion enhancement using local contrast method and geometrical and statistical feature extraction.	92.3	1	2	KNN
Huda et al. [23]	The normalization, data standardization and feature selection using feature importance property.	63	92	97	...	6	KNN
Reddy et al. [24]	The grid search parameter tuning optimization method and voting mechanism using ensemble based machine learning model.	65	2	KNN
Aabulwahhab et al. [25]	Discriminative interpretable features using socio-demographic and clinical information.	74	2	KNN
Proposed Method	Combined enhanced green & value color plane with PCA for feature extraction.	97.11	98.47	91.64	1	16	k-NN
		95.77	93.66	97.43	1	16	SVM

V. CONCLUSION

The early diagnosis of DR is a critical step in avoiding total blindness. The goal of our proposed method is to create an automated DR prediction system using Combined Enhanced Green and Value Planes (CEGVP) with k-nearest neighbor (k-NN) classifier in retinal fundus images to classify illnesses either individually or in combination. This was the first time that a 16 class classification was proposed that precisely gives the ability and flexibility to map the combinational complexity in a single step. The proposed work removes the requirement of four different binary classifiers for each abnormality detection and thus saves a lot of computational time, as well as error propagation from each method is also avoided. Our novel technique of 16 class classification using k-NN classifier achieved an accuracy of 97.11%, sensitivity of 98.47%, and specificity of 91.64% on DIARETDB0, DIARETDB1 & IDRid datasets. This new technique can help ophthalmologists discover problems more quickly and begin their diagnosis sooner. Future work includes improving the accuracy of the proposed system by using different machine learning classifiers. Besides RGB and HSV color planes techniques, other color planes can be explored and the useful plane can be merged to extract better features for DR detection.

ACKNOWLEDGMENT

The authors would like to thank colleagues from Sir Padampat Singhania University.

REFERENCES

- [1] Zimmet P, Alberti K, Shaw J. Global and societal implications of the diabetes epidemic. *Nature*. 2001;414(6865):782–787.
- [2] MacGillivray T, Trucco E, Cameron J, et al. Retinal imaging as a source of biomarkers for diagnosis, characterization and prognosis of chronic illness or long-term conditions. *The British journal of radiology*. 2014;87(1040):20130832.
- [3] Silvia RC, Vijayalakshmi R. Detection of non-proliferative diabetic retinopathy in fundus images of the human retina. In: 2013 International Conference on Information Communication and Embedded Systems (ICICES); IEEE; 2013. p. 978–983.
- [4] Coscas G, Cunha-Vaz J, Soubrane G. Macular edema: definition and basic concepts. *Macular Edema*. 2017;58:1–10.
- [5] Sreng S, Maneerat N, Hamamoto K, et al. Cotton wool spots detection in diabetic retinopathy based on adaptive thresholding and ant colony optimization coupling support vector machine. *IEEJ Transactions on Electrical and Electronic Engineering*. 2019; 14(6):884–893.
- [6] Arrigo A, Teussink M, Aragona E, et al. Multicolor imaging to detect different subtypes of retinal microaneurysms in diabetic retinopathy. *Eye*. 2021;35(1):277–281.
- [7] Yang D, Cao D, Huang Z, et al. Macular capillary perfusion in chinese patients with diabetic retinopathy obtained with optical coherence tomography angiography. *Ophthalmic Surgery, Lasers and Imaging Retina*. 2019;50(4):e88–e95.
- [8] Sisodia DS, Nair S, Khobragade P. Diabetic retinal fundus images: Preprocessing and feature extraction for early detection of diabetic retinopathy. *Biomedical and Pharmacology Journal*. 2017;10(2):615–626.
- [9] Nawaz IM, Rezzola S, Cancarini A, et al. Human vitreous in proliferative diabetic retinopathy: characterization and translational implications. *Progress in retinal and eye research*. 2019;72:100756.
- [10] Barsegian A, Kotlyar B, Lee J, et al. Diabetic retinopathy: focus on minority populations. *International journal of clinical endocrinology and metabolism*. 2017;3(1):034.
- [11] Campos GFC, Mastelini SM, Aguiar GJ, et al. Machine learning hyperparameter selection for contrast limited adaptive histogram equalization. *EURASIP Journal on Image and Video Processing*. 2019;2019(1):1–18.
- [12] Goyal B, Dogra A, Agrawal S, et al. Two-dimensional gray scale image denoising via morphological operations in nsst domain & bitonic filtering. *Future Generation Computer Systems*. 2018;82:158–175.
- [13] Lachure J, Deorankar A, Lachure S, et al. Diabetic retinopathy using morphological operations and machine learning. In: 2015 IEEE international advance computing conference (IACC); IEEE; 2015. p. 617–622.14.
- [14] Safitri DW, Juniati D. Classification of diabetic retinopathy using fractal dimension analysis of eye fundus image. In: AIP conference proceedings; Vol. 1867; AIP Publishing LLC; 2017. p. 020011.
- [15] Labhade JD, Chouthmol L, Deshmukh S. Diabetic retinopathy detection using soft computing techniques. In: 2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT); IEEE; 2016. p. 175–178.
- [16] Kushol R, Kabir MH, Salekin MS, et al. Contrast enhancement by top-hat and bottom-hat transform with optimal structuring element: Application to retinal vessel segmentation. In: International Conference Image Analysis and Recognition; Springer; 2017. p. 533–540.
- [17] Kaur J, Mittal D. A generalized method for the segmentation of exudates from pathological retinal fundus images. *Biocybernetics and Biomedical Engineering*. 2018;38(1):27–53.
- [18] Marin D, Gegundez-Arias ME, Ponte B, et al. An exudate detection method for diagnosis risk of diabetic macular edema in retinal images using feature-based and supervised classification. *Medical & biological engineering & computing*. 2018;56(8):1379–1390.
- [19] Issac A, Dutta MK, Travieso CM. Automatic computer vision-based detection and quantitative analysis of indicative parameters for grading of diabetic retinopathy. *Neural Computing and Applications*. 2020;32(20):15687–15697.
- [20] Chetoui M, Akhloufi MA, Kardouchi M. Diabetic retinopathy detection using machine learning and texture features. In: 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE); IEEE; 2018. p. 1–4.
- [21] Amin J, Sharif M, Rehman A, et al. Diabetic retinopathy detection and classification using hybrid feature set. *Microscopy research and technique*. 2018;81(9):990–996.
- [22] Sahu S, Singh AK, Ghrera S, et al. An approach for de-noising and contrast enhancement of retinal fundus image using clahe. *Optics & Laser Technology*. 2019;110:87–98.
- [23] Huda SA, Ila IJ, Sarder S, et al. An improved approach for detection of diabetic retinopathy using feature importance and machine learning algorithms. In: 2019 7th International Conference on Smart Computing & Communications (ICSCC); IEEE; 2019. p. 1–5.
- [24] Reddy GT, Bhattacharya S, Ramakrishnan SS, et al. An ensemble based machine learning model for diabetic retinopathy classification. In: 2020 international conference on emerging trends in information technology and engineering (ic-ETITE); IEEE; 2020. p. 1–6.
- [25] Alabdulwahhab K, Sami W, Mehmood T, et al. Automated detection of diabetic retinopathy using machine learning classifiers. *European Review for Medical and Pharmacological Sciences*. 2021;25(2):583–590.
- [26] Sharma A, Shinde S, Shaikh II, et al. Machine learning approach for detection of diabetic retinopathy with improved pre-processing. In: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS); IEEE; 2021. p. 517–522.
- [27] Kauppi, T., Kalesnykiene, V., Kamarainen, J.-K., Lensu, L., Sorri, I., Uusitalo, H., Kälviäinen, H., Pietilä, J., DIARETDB0: Evaluation Database and Methodology for Diabetic Retinopathy Algorithms, Technical report.

- [28] Kauppi, T., Kalesnykiene, V., Kamarainen, J.-K., Lensu, L., Sorri, I., Raninen A., Voutilainen R., Uusitalo, H., Kälviäinen, H., Pietilä, J., DIARETDB1 diabetic retinopathy database and evaluation protocol, Technical report.
- [29] Porwal, Prasanna, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabudde, and Fabrice Meriaudeau. "Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research." *Data* 3, no. 3 (2018): 25.

A Novel Secure Transposition Cipher Technique using Arbitrary Zigzag Patterns

Basil Al-Kasasbeh

Faculty of Computer Studies, Arab Open University (AOU), Riyadh, Kingdom of Saudi Arabia

Abstract—Symmetric cipher cryptography is an efficient technique for encrypting bits and letters to maintain secure communication and data confidentiality. Compared to asymmetric cipher cryptography, symmetric cipher has the speed advantage required for various real-time applications. Yet, with the distribution of micro-devices and the wider utilization of the Internet of Things (IoT) and Wireless Sensor Network (WSN), lightweight algorithms are required to operate on such devices. This paper proposes a symmetric cipher based on a scheme consisting of multiple zigzag patterns, a secret key of variable length, and block data of variable size. The proposed system uses transposition principles to generate various encryption patterns with a particular initial point over a grid. The total number of cells in the grid and its dimension are variable. Various patterns can be created for the same grid, leading to different outcomes on different grids. For a grid of n cells, a total of $n! * (n-1)!$ total patterns can be generated. This information is encapsulated in the private key. Thus, the huge number of possible patterns and the variation of the grid size, which are kept hidden, maintain the security of the proposed technique. Moreover, variable padding can be used; two paddings with different lengths lead to a completely different output even with the same pattern and the same inputs, which improves the security of the proposed system.

Keywords—Cryptography; symmetric cipher; block cipher; transposition algorithms

I. INTRODUCTION

Data confidentiality can be ensured using both asymmetric cipher and symmetric cipher. Symmetric cipher has the advantage of low computational requirements compared to asymmetric cipher cryptography, which is commonly used for key exchange and authentication purposes [1]. Asymmetric cipher is commonly referred to as public-key encryption [2]. On the other hand, the symmetric cipher is commonly referred to as private-key encryption and has two main types: the block cipher and the stream cipher. The stream cipher encrypts each bit of the message with each key's bit using operations, such as the exclusive-or (XOR) [3].

On the other hand, block cipher uses substitutions and transposition operations on an n -bits block, as illustrated in Fig. 1. Block cipher encrypts a data block of predetermined length using private key principles. Both stream and block ciphers have their applications, advantages, and disadvantages. The stream cipher is fast and requires low computational power compared to the block cipher. Block cipher is widespread and is used in various other applications, such as stream cipher, hash function, pseudorandom number generator, and message authentication [4]. Besides, the block cipher is more secure than the stream cipher, subject to the strength of the utilized

private key [5]. The security level of the block cipher algorithm is evaluated based on the complexity, the performance, and its strength against possible cryptanalyses, such as linear and differential analysis and the homegrown crypto that is created afterthought [4, 6, 7].

Block cipher algorithms use predefined block sizes. A multiple of 8-bits block sizes is commonly utilized as compatible with most processors. Besides, the large size is avoided as it leads to padding, increasing computational requirements. Similarly, small size blocks give more chances for dictionary attacks on the ciphertext blocks to succeed [8]. For a variable-length message to be encrypted, the message is first divided into blocks of the predetermined size, padded if necessary to meet the required size. Then each block is encrypted using the cipher algorithm with a private key. Advanced algorithms for block cipher have been developed, such as Advanced Encryption Standard (AES) and Data Encryption Standard (DES), which are the most utilized encryption algorithms worldwide [4, 9]. These algorithms depend on mathematical operation and substituting the plaintext bits and characters by other bits and characters; these are called substitutional algorithms. Besides the substitution cipher, another block cipher type is called transposition cipher, such as rail fence and columnar. Compared to the substitution cipher, the transposition cipher is faster, as it depends on transposition shuffle rather than mathematical operations. Yet, because the shuffle is limited and known in advance, and the algorithms are operated on very small keys, the security of these algorithms is weak. Surprisingly, vast number of the existing substitutional algorithms has been solved by the cryptanalysis [4, 5, 10, 11].

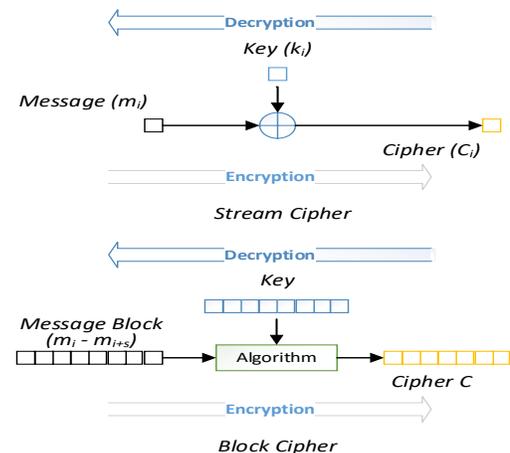


Fig. 1. Stream Cipher vs. Block Cipher.

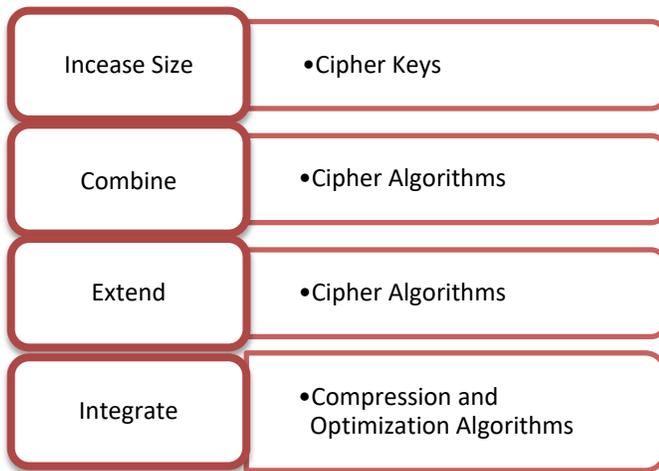


Fig. 2. Modern Cryptograph Directions.

Accordingly, various modern cipher techniques have been proposed to improve the security of the block cipher. Various ways have been followed to extend the existing algorithms. As illustrated in Fig. 2, these are 1) Proposing various encryption modes, in which the block ciphers are related to each other or related to randomly created vectors to increase the complexity of the generated cipher. 2) Combining various cipher algorithms to create a more secure system. 3) Combining cipher algorithms with optimization and compression algorithms. 4) Extending the cipher algorithms by increasing the size of the key or modifying their structure. Although these directions improve the security of the cipher algorithms, most of them have increased the computational requirements of the encryption and decryption process. The complexity of the modern techniques for data encryption leads to the inapplicability of these techniques to be utilized in most of the modern applications, such as real-time applications, sensor networks [12], and Internet-of-Things (IoT) [13], which demands secure yet highly performance encryption techniques [14].

Accordingly, this paper proposes a new technique for block cipher encryption. The proposed system follows different directions compared to these mentioned previously in improving the security of the symmetric cipher cryptography [15]. The goal of the proposed technique and the creative direction is to improve security while maintaining low encryption and decryption processes requirements. The proposed technique improves the security using different transposition zigzag patterns while solving one of the major transposition cipher problems: the limited possible shuffle operation of the cipher algorithm. Moreover, the proposed technique used variable grid size maintained secret in the private key, similar to the columnar algorithm. Thus, before the encryption is made available for the sender, various patterns are created and saved with their identification number in the system. These patterns are shared publically between the sender and the receiver. However, trying all these patterns is impossible for cryptanalysis because of the vast possible patterns. Then, the proposed system allows the sender to freely choose among different patterns and select the grid size on

which the patterns are drawn. The size and the pattern form the private key of this cipher system. A variable-length message can be divided into blocks of various lengths, padded if necessary to meet the required length. Each block is encrypted using different patterns and keys.

II. LITERATURE REVIEW

Transposition ciphers are implemented based on designed patterns to scramble the plaintext. Examples of some of the transposition ciphers are presented in Fig. 3. The rail fence used the number of rows as the key. Then, a grid is created with the specified number of rows and number of columns equal to the length of the message to be encrypted. The plain text is placed diagonally downwards on successive rails on an imaginary fence, followed by upwards moves as the movement reaches the grid's last row. The letters are read row by row to create the ciphertext [16]. In the columnar cipher, the key determines the number of columns and the columns' order. Then, a grid is created with a specific number of columns and rows to accommodate all the letters in the message to be encrypted. The plain text is placed on the grid row by row and in order. Then, the columns are permuted based on the key. The letters are read column by column to create the ciphertext [16-19].

The route cipher is another grid-based transposition cipher with longer keys than the rail fence and columnar. The key represents the number of columns or rows and the patterns to read the letters in the grid. The plaintext is written in the grid row by row and in order. Then, the ciphertext is produced by reading the grid upwards or downwards clockwise or zigzag patterns up and down. Double or multiple columnar ciphers are used to improve the security of the single columnar cipher. The same key can be applied, or other keys are required in multiple rounds of the columnar cipher process [18]. Myszkowski is a variation of the columnar cipher with a key of repeated letters that are given the same order. Accordingly, columns with unique numbers are read downward, and columns with recurring numbers are transcribed left to right. Disrupted transposition is another grid-based cipher with irregular spaces added between the plaintext letters. Overall, various transposition ciphers have been proposed. These transposition ciphers depend on a grid to scramble the plaintext. Yet, the problem of these ciphers is the limited patterns with a small key space that can be discovered in the cryptanalysis processes that are searching the keyspace [20].

Various modern techniques were proposed by modifying the transposition cipher algorithm. Sokouti, Sokouti [21] added 8-bits padding to each 8-bits in the message, structured these bits in a binary tree, and traversed this tree using an in-order traversal algorithm. However, such complex processes increase the complexity of the transposition cipher. Twum, Hayfron-Acquah [22] extended columnar cipher with a modified Rubik's cube puzzle with a higher level of security as the cube represents a 3-dimensional instead of the 2-dimensional space utilized in the original columnar cipher. The number of such modified algorithms is enormous. Block cipher surveys were presented by Surya and Diviya [23], Albermany and Radihamade [10], and Mandal [24].

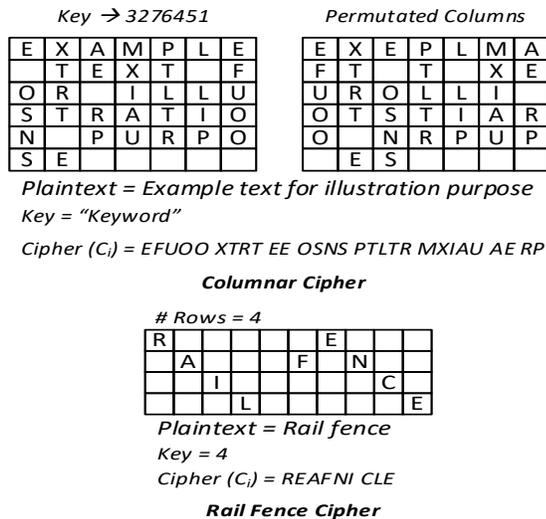


Fig. 3. Example Transposition Ciphers.

Lasry, Kopal [18] extended columnar cipher with a two-phases hill-climbing algorithm in combining cipher algorithms with optimization. The main goal of the developed algorithm is to increase the size of the key to improve the security of the cipher. Hill climbing is an optimization algorithm that starts with a random solution and iteratively finds optimal. With columnar cipher, hill climbing is used to swap columns and increase the adjacency score of the resulting cipher. Various other schemes combined cipher algorithms with optimization algorithms, such as genetic algorithm (GA) [25, 26]. Using the Fibonacci code algorithm, Siregar, Fadlina [19] integrated columnar cipher and data compression. The columnar output is input to the Fibonacci algorithm, which produces the final ciphertext.

Encryption modes describe how the blocks of a single message are related to each other to improve the security of the message. The electronic codebook (ECB) encryption mode for block cipher encrypts each block individually. The advantages of this mode are the ability of out-of-order decrypting, non-propagation of errors between blocks, speed, and parallelism. Yet, as the same key is utilized, the same input plaintext produces the same cipher, which is deterministic and can be attacked through traffic analysis [4]. To overcome this limitation, several encryption modes have been developed; these are cipher block chaining (CBC), cipher feedback (CFB), output feedback (OFB), and counter (CTR) mode [11]. These modes operated based on the concept of using extra data besides plaintext and the key for creating probabilistic encryption. CBC encrypts the first block together with a randomly initialized vector. The vector is combined with plain text before the encryption process. The ciphertext is then used with the second block in place of the random vector, and the output is used with the third block and so on. In CFB, the random vector is encrypted and combined with plaintext, which emulates the self-synchronizing stream cipher. OFB is similar to CFB except that the encrypted vector is sent to the next block instead of the ciphertext, similar to the synchronous stream cipher concept. CTR mode initializes a unique vector for each block, encrypts the vector then combines the output

with the plaintext to produce the ciphertext [27]. These modes aim to scramble the output, avoid repeated patterns, and provide semantic security, to prevent inference of any information from the ciphertext under an unknown key [4, 11]. Yet, these modes create various disadvantages, mainly related to the computational power and the required processing time. Generally, these modes have been developed with substitution encryption in mind. Although these modes applied for transposition, less attention has been given to this encryption scheme, mainly because of the limitation related to its weak security, resulting from the weakness of their keys, and the limited variation of the transposition shuffle [28-30].

In the integration direction, Srikantaswamy and Phaneendra [31] integrated columnar transposition cipher with Caesar substitution cipher. The secret key is generated randomly based on a selected seed value. Then, Caesar cipher is implemented with alphabets, symbols, and numbers as an extension to the original version that uses alphabets only. Finally, columnar is implemented using a randomly selected number of columns. Kester [17] integrated columnar transposition cipher with Vigenere substitution cipher. Columnar is implemented using a randomly selected number of columns. The generated ciphertext is then used as the key for encrypting the plaintext using the Vigenere cipher. Dar [16] integrated columnar cipher, Caesar cipher, and rail fence cipher to improve the security of the ciphertext. The aforementioned cipher algorithms are implemented in order; as such, the output of the columnar is used as input to Caesar, and the output of Caesar is used as input to the rail fence, which produces the final ciphertext.

As noted, these extended techniques rely on more processing rounds and complex calculation and transposition to secure the data. Yet, those techniques cannot be applied for micro-devices with limited resources or real-time applications requiring rapid encryption-decryption processes [32]. Accordingly, this research proposed a strong security technique resulting from the possible variation with low processing requirements.

III. THE PROPOSED SYSTEM

The proposed technique relies on predefined zigzag patterns drawn on a two-dimensional grid and can be extended to three-dimensional as required. The patterns can be created manually, as illustrated in Fig. 4, or automatically using graph-based search, or randomly depending on the application and the device. In a typical implementation of the proposed techniques, these patterns are made publically available and exchanged between the sender and the receiver. Yet, to make the system secure, the number of these patterns should be large enough to avoid brute-force solving of the ciphertext. Moreover, some patterns can be made secret and shared only between the sender and the receiver.

Besides, these patterns are saved in the sender and receiver devices in two arrays, one for the x-axis and the other for the y-axis, as illustrated in Fig. 5. If a three-dimensional cube is utilized, another array is used for the z-axis. Such representation requires low memory and can be traversed with O(n) complexity, where n is the size of the grid. The vast number of possible patterns maintains security. For a grid of n

cells, a total of $(n!)$ can be generated. Moreover, given that n is variable, $(n! * (n-1!))$ total patterns can be generated if n is hidden. The value of n can be made invisible as the proposed technique allows for arbitrary padding at the end of the message before and after encryption. For example, for a message of length 6, 720 $(6!)$ patterns are drawn on a grid of size 6 $(2*3)$. Yet, after encryption padding, the message can appear of size 9, for example. Thus, for cryptanalysis to try all possible patterns, all patterns of the grid of sizes $\{9, 8, \dots, 1\}$ should be examined.

The line-crossing patterns can be created manually but preferably using random number generation. The grid size and the initial cell s are determined to generate a manual pattern. Then, the pattern is created following zigzag directions visiting all the cells once and terminating the traversing at the arbitrary final cell, e . The coordinates of the cells are preserved in the arrays based on the traversing order. Each pattern is saved with a unique identifier to reference between the sender and the receiver. Besides, a generated pattern can be kept secret and exchanged with the private key because the pattern can be represented as two arrays in a two-dimensional grid. Although the zigzag patterns required manual attention to avoid line-crossing, these zigzag patterns can be generated using a graph representation of the grid and graph traversals mechanisms, such as in-order, pre-order, and post-order, as illustrated in Fig. 6. Nevertheless, as illustrated in Fig. 7, line-crossing patterns do not affect the security of the proposed technique. Instead, such line-crossing patterns allow for random patterns without the need for human attention or graph representation and traversals.

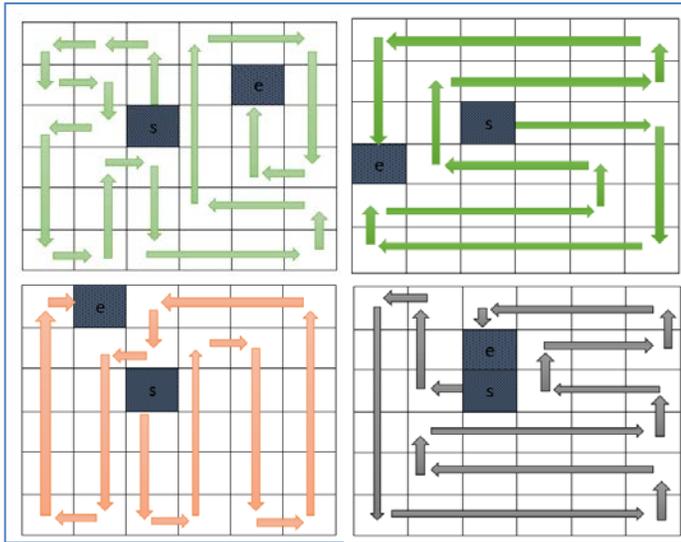


Fig. 4. Zigzag Patterns Examples.

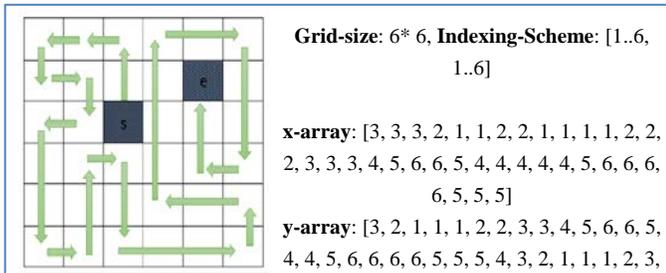


Fig. 5. Example of Array Representation of a Zigzag Pattern.

A. Pattern Generation

The pattern by which the transposition processes are implemented can be created in two forms and three ways. The zigzag patterns, which do not include any line-crossing, can be generated either manually or using graph-based techniques, as listed in Table I.

TABLE I. PATTERNS LIST OF THE PROPOSED TECHNIQUE

Pattern Type	Generated Mechanism	Advantages
Non-crossing Zigzag	Manual	Visually interpreted and verifiable.
Returnable Pattern	Graph Traversal (In-order, pre-order, and post-order)	Visually interpreted and verifiable and created automatically
Crossing-Line Pattern	Manual or Randomly	Easily created with wide varieties

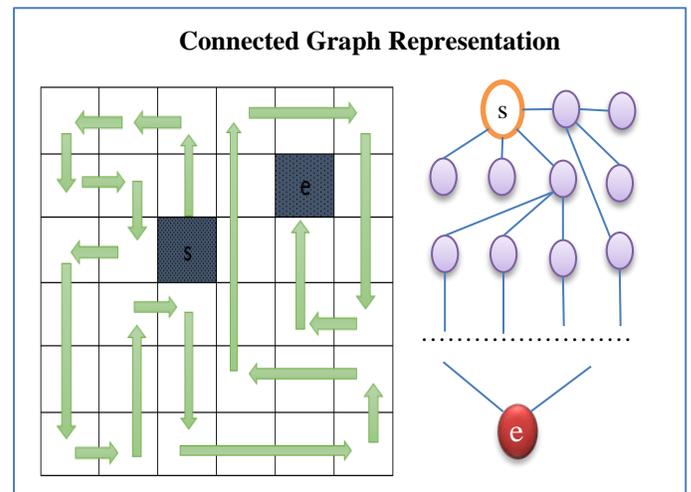


Fig. 6. The Graph Representation for a Zigzag Pattern Generation.

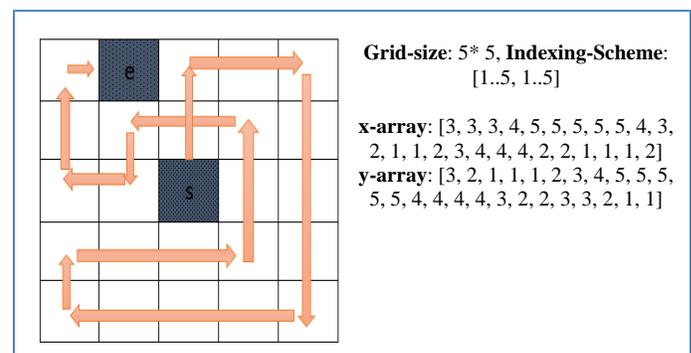


Fig. 7. Crossing-Line Pattern.

B. Secret Key Value

The secret or the private key of the proposed technique is represented with an integer value, which can be converted into binary as required. The private key of the proposed cipher technique referred to the grid size, the pattern to be utilized, and the padding options. There are two options for determining

the pattern, either a secret pattern embedded into the private key or the unique identifier of a public pattern. A secret pattern can be identified as starting by -1 value. The public pattern is identified with a unique positive identifier. The identifier can be of any length to continuously add patterns without limits. There are also two padding options to improve the security of the proposed technique, as listed in Table II. Table II lists the private key components in the proposed technique with examples.

TABLE II. PRIVATE KEY COMPONENTS OF THE PROPOSED TECHNIQUE

Components		Description	Example
Length		Four digits: The first digit indicates using public (0) or secret pattern (1). The next two digits indicated the length of the pattern (either length of the identifier or length of the included pattern). The next digit indicates whether padding is done before (0) or after encoding (1). If no padding occurs, then this digit can take any value.	0160 → using public pattern identified by 16 digits with padding implemented before encrypting
Grid size		4-digits for the two-dimensional	0610 → six rows and ten columns
Pattern	Public Identifier	Variable number of digits for the unique identifier	Can be any number
	Secret Pattern	Two concatenated arrays of size equal to the grid size (identified by the size field in the pattern)	21211212 → two concatenated arrays [2121] and [1212] that describe a pattern over 2*2 grid
Padding	Before encoding	Two digits to indicate the length of the padding at the end of the grid	05 → padding of length 5
	After encoding	Two digits to indicate the length of the padding at the end of the ciphertext	04 → padding of length 4
Example Keys			
1501 0505 3334555554321123444221112321112345555444432233211 06			
A secret pattern consists of 50 digits of the secret pattern or a 5x5 grid with six digits padding before encrypting			
0101 0505 0125214578 06			
A public pattern has an identity consisting of 10 digits with the same specification as the previous one			

C. Encryption

The encryption process is implemented based on the specifications stated in the secret key. First, the grid is created based on the predetermined dimensions, utilizing the secret key components. The plain text components are placed inside the grid based on the pattern represented by the x-array and the y-array. The ciphertext is produced by reading the grid row by row as the grid is filled. If padding is implemented before encoding, then the padded code (either 0's in case of encrypting binaries or X's in the case of encrypting letters) is considered in the last row(s) of the created grid. For example,

if a 3x3 grid is utilized with two padding bits before encoding, then the value of the cells $\{(2,3), (3,3)\}$, will be zeros, and these cells will not be filled with the components of the plaintext as will be explained. If padding is implemented after the encoding, the padded code is attached after the ciphertext is produced. Algorithm 1 presents the encryption process. The significance of the padding is that, even with the same pattern and the same inputs, different pad lengths for the before encoding padding leads to a completely different output.

Algorithm 1: Encryption

```

1 Input: mb, K (cols, rows, xs [], ys [], beforePadding, afterPadding)
2 Output: cb
3 Grid [[]]:= Create-Grid (cols, rows)
4 xPadding [],yPadding []:= IdentifyPaddingCells(beforePadding)
5 For-Each x,y in xs, ys
6     IF (x,y ∈ xPadding , yPadding)
7         Grid [x,y] := 0
8     ELSE
9         Grid [x,y] := mbi, i++
10 For-Each c in Grid
11     cbi := c
12 IF (afterPadding > 0)
13     Concatenate (cb, pad)
    
```

Notations:

mb, cb: message block and the cipher block, respectively
k: private key
cols, rows: number of columns, and rows as determined by the private key, respectively
xs, ys: the x-array and y-array of the pattern, respectively.
beforePadding, afterPadding: length of the padding before and after encryption, respectively.

As given in Algorithm 1, line 3 created the grid used by the encryption. The padded cells used before encryption are identified based on their numbers, as the last cells in the grid, as given in line 4. Lines 5-9 fill the grid following the pattern with the message (lines 8-9) or with padding (lines 6-7) if the pattern comes across the padding cell. Lines 10-11 create the ciphertext by reading the grid row by row. Finally, lines 12-13 concatenate after encryption pads if it exists as determined in the key.

D. Decryption

The grid is created based on the predetermined dimensions after receiving the encrypted message (bit sequence or letter sequence) at the receiver side. The grid is filled in a row by row manner, similar to how the sender created the ciphertext in the last stage. Then, to retrieve the plaintext, the receiver will implement the same pattern as utilized in the receiver side, but this time to read the cells and reproduce the message. The padded code cells are skipped while reading the grid components if padding is implemented before encoding. At the same time, if padding is implemented after encoding, the last components of the received ciphertext will be removed. Algorithm 2 presents the decryption process.

Algorithm 2 represents opposite operations. Line 3 created the grid. After encryption pads are removed, then in lines 5-6. Then, the grid is filled in lines 7-8. Lines 5-9 read the grid following the pattern and filling the message (lines 12-13) or skip padding (lines 10-11) if the pattern crosses the padding cell.

Algorithm 2: Decryption

```

1  Input: cb, K (cols, rows, xs [], ys [], beforePadding, afterPadding)
2  Output: mb
3  Grid [[]]:= Create-Grid (cols, rows)
4  xPadding [],yPadding []:= IdentifyPaddingCells(beforePadding)
5  IF (afterPadding > 0)
6    Eliminate (cb, pad)
7  For-Each c in Grid
8    cbi := c
9  For-Each x,y in xs, ys
10   IF (x,y ∈ xPadding , yPadding)
11     continue
12   ELSE
13     mbi := Grid [x,y], i++
    
```

Notations:

mb, cb: message block and the cipher block, respectively
k: private key
cols, rows: number of columns, and rows as determined by the private key, respectively
xs, ys: the x-array and y-array of the pattern, respectively.
beforePadding, afterPadding: length of the padding before and after encryption, respectively.

E. Multiple Encryption-Decryption

Multiple patterns can be used with the same block with multiple rounds while considering the computational requirements at the sender and the receiver, as given in Fig. 8. Accordingly, multiple private keys are required for such a purpose. Nevertheless, multiple rounds can be done with the same key, as each round will contribute to the scrambling of the input.

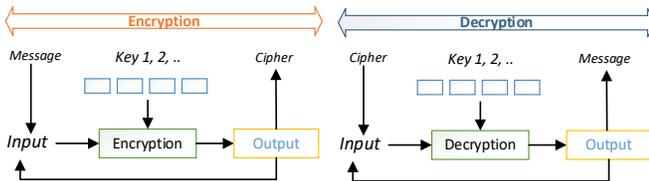
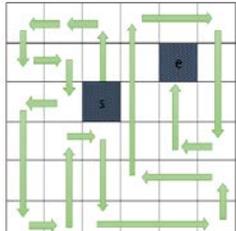


Fig. 8. Multiple Encryption-decryption.

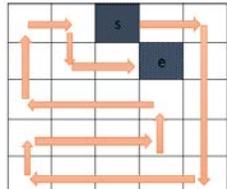
IV. EXAMPLES

An example of encrypting and decrypting a plaintext represented using a binary code and a private key, which specify the size of the grid and the pattern, is given in Fig. 9. A single block is encrypted and decrypted using the given key in the example. The rest of the blocks can be encrypted and decrypted in the same way if the same key is utilized. Otherwise, different grid-size and different patterns are utilized.



Key: 0101|0606|2025214515|00
Message:
 100111001011010001010001011110100111
Cipher:
 110110100111101110000010110100101010

Fig. 9. Example of Binary Encrypting and Decrypting.

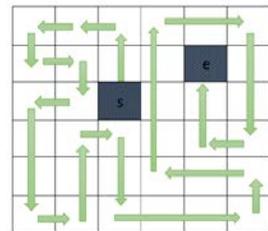


Key: 0091|0505|145786952|00
Message:
 ABCDEFGHIJKLMNOPQRSTUVWXYZ
Cipher:
 VWABCUXYZDTSRQELMNPEKJIHG

Fig. 10. Example of Letters Encrypting and Decrypting .

As shown in Fig. 10, the second example represents the encrypting and decrypting of letters and numbers. The process is identical to the previous examples. Using different keys leads to using different grid sizes, different patterns, and different encoding padding. The letters in the example can be encrypted directly as given in the example or converted into binary code and encrypted bit-wise.

As given in Fig. 11, the third example represents the encrypting and decrypting of binary code before encoding padding. As noted, the results were contributing to more scrambling of the message.



Key: 0100|0606|2025214515|11
Message:
 ABCDEFGHIJKLMNOPQRSTUVWXYZ
Cipher:
 EDCRSTFGBQZUIHAPYVKLMOXWK

Fig. 11. Example of Binary Encrypting and Decrypting with Padding.

V. SECURITY CONCERNS

Unlike the substitution cipher, attacks on transposition cipher do not depend on linear and differential cryptanalysis [33, 34]. Instead, the attacks on transposition cipher depend on guessing the key with statistical analysis of the n-gram of the language [35]. In such a process, the optimization algorithms reduce the time required by the brute-force approach. Accordingly, there are two strength issues to defeat such cryptanalysis: increasing the keyspace and increasing the possibilities of the transposition and scrambling processes. In the proposed technique, the key size has been increased to variable size, given that the size can be encapsulated into the private key itself. Although the size of the block can reveal much about the grid possibilities, using padding in two ways increases the block's size and leads to variable block size. Accordingly, padding the text with long pads are preferable to increase the security of the proposed technique. Moreover, as the encryption patterns are hidden in the private key, it is hard to implement all possible patterns to discover the correct one. Semantic security is granted in the proposed technique as it cannot infer any information from the ciphertext under an unknown key. To summarize, the security of the proposed technique can be listed as given in Table III.

The proposed system's validation is implemented based on various texts encrypted with random keys and with reference to the frequency distribution graph of the standard English letter, as illustrated in Fig. 12. This validation assumes that the

attacker captures the ciphertext without knowing the key. Accordingly, the attacker implements frequency distribution to analyze the captured ciphertext. An example of this graph is given in Fig. 13(a). The attacker then compares the frequency distribution graph with the Standard English Letter. Matching is then implemented by shifting the distribution graph to match the distribution graph of the English letters. As given in Fig. 13(b), the matching can be found by shifting the ciphertext graph by 3 letters. Shifting has been determined as the letter "E" is the most frequent in English, while letter H is the most frequent letter of the ciphertext graph, as it has been used in padding. Yet, the plaintext was not obtained as the reverse shifting is implemented. Accordingly, frequency analysis of 1000 different samples and shifting is determined by matching the frequency graph of each sample with the letter "E". The obtained results after shifting did not match any input text of these samples.

TABLE III. STRENGTH ASPECTS OF THE PROPOSED TECHNIQUE

Strength Aspect	Description
Vast patterns	The number of patterns that can be created is large
Variable grid size	Using padding, the actual size of the grid is hidden
Long key	The key length can be large with patterns included or long pattern identity
Hidden pattern	Secret patterns make the process of breaking the cipher harder

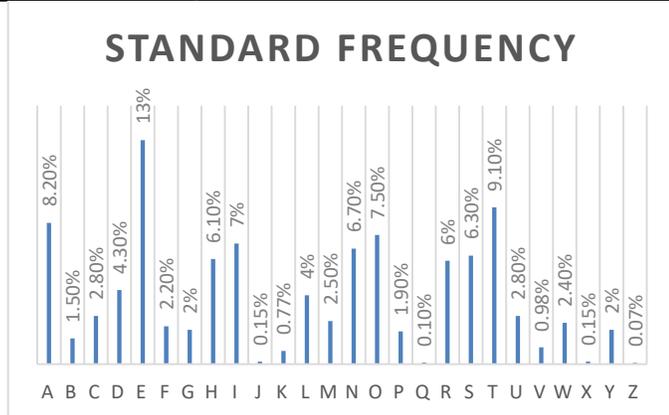
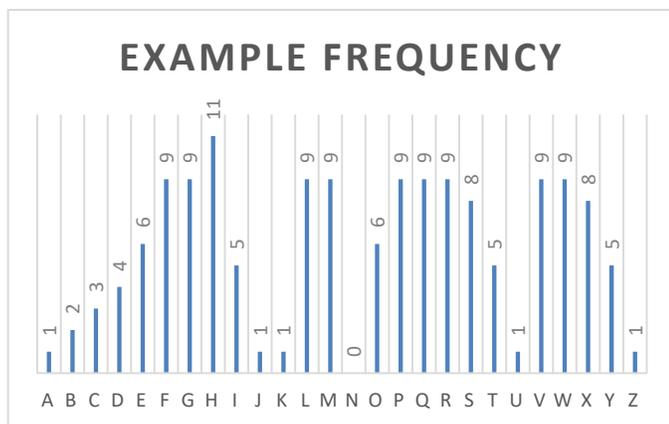
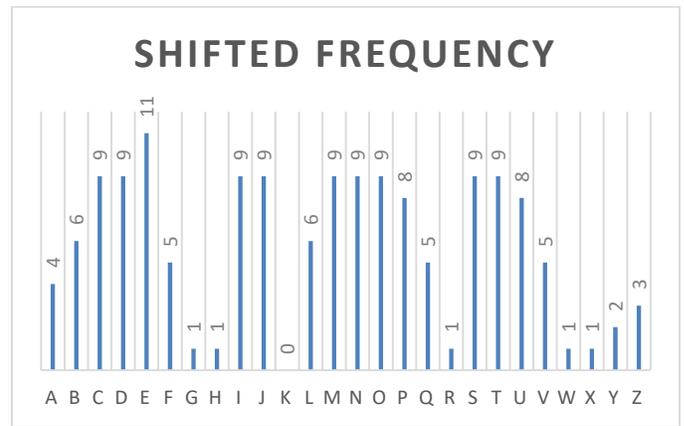


Fig. 12. Standard English Text Frequency.



(a) Example Frequency Graph.



(b) Example Frequency Graph after Shifting.

Fig. 13. Frequency Analysis of Ciphertext.

VI. CONCLUSION

In this paper, a technique for a transposition block cipher is proposed based on arbitrary grid size, initial point, and arbitrary zigzag patterns. The proposed technique improves the security of the ciphertext by scrambling the letters or the bits of the plaintext in an unpredicted manner. The time and memory requirements of the proposed technique are maintained as low as possible to cope with the requirements of currently utilized micro-devices and real-time applications. Accordingly, the proposed technique is of $O(n)$ complexity, where n is the grid's size. High security is maintained using a large key, data block of variable size with vast possible patterns. To defeat the cryptanalysis, the block size is increased using padding in two ways. Two similar paddings, yet with different lengths, lead to a completely different output even with the same pattern and the same inputs, which improves the security of the proposed system. Accordingly, hiding the patterns, the block size, and the padding in the private key makes it hard to implement all possible patterns to discover the correct one. The frequency analysis implemented proves the security of the proposed technique.

ACKNOWLEDGMENT

The author would like to thank Arab Open University-KSA and Oracle-KSA for supporting this study.

REFERENCES

- [1] N. Li, "Research on Diffie-Hellman key exchange protocol," 2nd International Conference on Computer Engineering and Technology, Chengdu, China, 2010, pp. 634-637.
- [2] M.R. Joshi and R.A. Karkade, "Network security with cryptography," International Journal of Computer Science and Mobile Computing, vol. 4(1), 2015, pp. 201-204.
- [3] M. U. Bokhari, S. Alam, and F.S. Masoodi, "Cryptanalysis techniques for stream cipher: a survey," International Journal of Computer Applications, vol. 60(9), 2012.
- [4] D. Bujari, and E. Aribas, "Comparative analysis of block cipher modes of operation," International Advanced Researches & Engineering Congress, Osmaniye, Turkey, 2017, pp. 1-4.
- [5] Valea, E., et al., Stream vs block ciphers for scan encryption. Microelectronics Journal, 2019. 86: p. 65-76.
- [6] S. Rani and H. Kaur, "Technical review on symmetric and asymmetric cryptography algorithms," International Journal of Advanced Research in Computer Science, vol. 8(4), 2017.

- [7] A.G. Khan, S. Basharat, and M.U. Riaz, "Analysis of asymmetric cryptography in information security based on computational study to ensure confidentiality during information exchange," *International Journal of Scientific & Engineering Research*, vol. 9(10), 2018, pp. 992-999.
- [8] G. Singh, A. K. Singla, and K. S. Sandha., "Superiority of blowfish algorithm in wireless networks," *International Journal of Computer Applications*, vol. 44(11), 2012, pp. 23-26.
- [9] S. Singh, S.K. Maakar, and S. Kumar, "A performance analysis of DES and RSA cryptography," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, vol. 2(3), 2013, pp. 418-423.
- [10] S. Albermany, and F. Radihamade, "Survey: block cipher methods," *Int. J. Adv. Res. Technol*, vol. 5(11), 2016, pp. 11-22.
- [11] P. Sharma and R. Purohit, "Performance evaluation of symmetric block cipher RC6 with ECB and CBC operation modes," *International Conference on Intelligent Data Communication Technologies and Internet of Things, Coimbatore, India, 2018*, pp. 134-140.
- [12] W. Mazurczyk, "VoIP steganography and its detection—a survey," *ACM Computing Surveys (CSUR)*, vol. 46(2), 2013, pp. 1-21.
- [13] P. Asghari, A.M. Rahmani, and H.H.S. Javadi, "Internet of Things applications: A systematic review," *Computer Networks*, vol. 148, 2019, pp. 241-261.
- [14] S. Wachter, "Normative challenges of identification in the Internet of Things: Privacy, profiling, discrimination, and the GDPR," *Computer law & security review*, vol. 34(3), 2018, pp. 436-449.
- [15] M. A. Hossain, M. B. Hossain, M. S. Uddin, and S. M. Imtiaz, "Performance analysis of different cryptography algorithms," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 6(3), 2016, pp. 659-665.
- [16] J.A. Dar, "Enhancing the data security of simple columnar transposition cipher by caesar cipher and rail fence cipher technique," *International Journal of Computer Science & Engineering Technology (IJCSSET)*, vol. 5(11), 2014, pp. 1054-1061.
- [17] Q. A. Kester, "A hybrid cryptosystem based on vigenere cipher and columnar transposition cipher," *International Journal of Advanced Technology and Engineering Research (IJATER)*, vol. 3(11), 2013, pp. 141-147.
- [18] G. Lasry, N. Kopal, and A. Wacker, "Cryptanalysis of columnar transposition cipher with long keys," *Cryptologia*, vol. 40(4), 2016, pp. 374-398.
- [19] S. R. Siregar, F. Fadlina, and S.D. Nasution, "Enhancing data security of columnar transposition cipher by fibonacci codes algorithm," *Third Workshop on Multidisciplinary and Its Applications, WMA-3, Medan, Indonesia, 2019*, pp. 1-10.
- [20] A. Bhowmic, and M. Geetha, "Enhancing resistance of hill cipher using columnar and Myszkowski transposition," *International Journal of Computer Sciences and Engineering*, vol. 3(2), 2015, pp. 20-25.
- [21] M. Sokouti, B. Sokouti, and S. Pashazadeh, "An approach in improving transposition cipher system," *Indian Journal of Science and Technology*, vol. 2(8), 2009, pp. 9-15.
- [22] F. Twum, J. Hayfron-Acquah, and W. Morgan-Darko, "A proposed enhanced transposition cipher algorithm based on rubik's cube transformations," *International Journal of Computer Applications*, vol. 182(35), 2019, pp. 18-26.
- [23] E. Surya, and C. Diviya, "A survey on symmetric key encryption algorithms," *International Journal of Computer Science & Communication Networks*, vol. 2(4), 2012, pp. 475-477.
- [24] P.C. Mandal, "Evaluation of performance of the symmetric key algorithms: des, 3des, aes and blowfish". *Journal of Global Research in Computer Science*, vol. 3(8), 2012, pp. 67-70.
- [25] E. A. M. Eliáš, "Evolutionary computation in cryptanalysis of classical ciphers," *Tatra Mt. Math. Publ.*, vol. 70, 2017, pp. 179-197.
- [26] S. Picsek, and D. Jakobovic, "Evolutionary computation and machine learning in cryptography," *Genetic and Evolutionary Computation Conference Companion, Cancún, Mexico, 2020*, pp. 1147-1173.
- [27] K.K. Pandey, V. Rangari, and S. Kumar, "An enhanced symmetric key cryptography algorithm to improve data security," *International Journal of Computer Applications*, vol. 74, 2013, pp. 0975 – 8887.
- [28] O. Omolara, A. Oludare, and S. Abdulahi, "Developing a modified Hybrid Caesar cipher and Vigenere cipher for secure data communication," *Computer Engineering and Intelligent Systems*, vol. 5(5), 2014.
- [29] P. Singh, and P. Shende, "Symmetric key cryptography: current trends", *International Journal of Computer Science and Mobile Computing*, vol. 3(2), 2014, pp. 410-415.
- [30] X. Yi, R. Paulet, and E. Bertino, "Homomorphic encryption," *Homomorphic Encryption and Applications*, 2014, pp. 27-46.
- [31] S. Srikanthaswamy, S. and D.H. Phaneendra, "Improved Caesar cipher with random number generation technique and multistage encryption," *International Journal on Cryptography and Information Security (IJCIS)*, vol. 2(4), 2012, pp. 39-49.
- [32] R. Rejani, and D.V. Krishnan, "Study of symmetric key cryptography algorithms," *International Journal of Computer Techniques*, vol. 2(2), 2015, pp. 45-50.
- [33] L. Keliher, "Refined analysis of bounds related to linear and differential cryptanalysis for the AES," *International Conference on Advanced Encryption Standard, Bonn, Germany, 2004*, pp. 42-57.
- [34] L. R. Knudsen, "Block Ciphers—a survey," *State of the art in applied cryptography*, Berlin, Heidelberg, 1998, pp. 18-48.
- [35] A. M. Kadhim, "Diagnosis of some cipher systems. journal of baghdad college of economic sciences university," vol. 4, 2013.

Feature Concatenation based Multilayered Sparse Tensor for Debond Detection Optical Thermography

Junaid Ahmed¹, Abdul Baseer², Guiyun Tian³, Gulsher Baloch⁴, Ahmed Ali Shah⁵

Electrical Engineering Department, Sukkur IBA University, Sukkur, 65200, Sindh, Pakistan^{1,2,4,5}
School of Automation Engineering, University of Electronic Science and Technology, Chengdu, Sichuan, China³

Abstract—Composites being the key ingredients of the manufacturing in the aerospace, aircraft, civil and related industries, it is quite important to check its quality and health during its manufacture or in service. The most commonly found problem in the CFRPs is debonding. As debonds are subsurface defects, the general methods are not quite effective and require destructive tests. The Optical Pulse Thermography (OPT) is a quite promising technology that is being used for detecting the debonds. However, the thermographic time sequences from the OPT system have a lot of noise and normally the defects information is not clear. For solving this problem, an improved tensor nuclear norm (I-TNN) decomposition is proposed in the concatenated feature space with multilayer tensor decomposition. The proposed algorithm utilizes the frontal slice of the tensor to define the TNN and the core singular matrix is further decomposed to utilize the information in the third mode of the tensor. The concatenation helps embed the low-rank and sparse data jointly for weak defect extraction. To show the efficacy and robustness of the algorithm experiments are conducted and comparisons are presented with other algorithms.

Keywords—Improved tensor nuclear norm; low-rank decomposition; concatenated feature space; optical thermography

I. INTRODUCTION

For the task of extracting weak defect information in the thermal sequences of carbon fiber reinforced polymer (CFRP) debonds using the optical pulse thermography (OPT) based technology, post-image processing techniques are generally used. In [1], principal component analysis (PCA) is used along with the OPT for detecting the debonds in the CFRP. The PCA algorithms decompose the thermal sequence data into a low dimensional space using either the eigenvalue decomposition or the singular value decomposition (SVD). The algorithm provides reasonable results for detecting the debond defects in the CFRPs. In [2], another decomposition-based algorithm is proposed called the independent component analysis (ICA). The algorithm is similar to the PCA and provides reasonable results for detecting the debonds in CFRPs. In [3], the authors propose a polynomial-based decomposition algorithm called the thermal signal reconstruction (TSR). It works in the logarithmic domain and performs the polynomial fitting to extract the defect information in the thermal video sequences. The algorithm performs a little better than PCA and ICA however, it has a long-running time. In [4], [5], pulse phase thermography (PPT) is proposed. It employs an extension to the TSR algorithm in the frequency domain and extracts the amplitude and phase information for defect analysis. The PPT algorithm utilizes the Fourier Transform for extracting the

defects information in the thermal sequences. In [6], [7], sparse principal component thermography (SPCT) is proposed for debonding detection in composites. This algorithm is an extension of the PCA algorithm. It induces sparsity into the algorithm. This algorithm works well for the flat shape specimen. In [8], feature embedding is proposed which uses the joint feature space for low-rank and sparse analysis. In [9], a tensor decomposition algorithm is proposed called the ensemble variational Bayes tensor factorization (EVBTF). It employs a multilayer architecture with tensor decomposition. The algorithm is used for detecting the debonds in the CFRP specimen using optical pulse thermography. The specimen under test was the flat rectangular shape CFRPs with debond defects at multiple depths and with multiple diameters. Another multilayer decomposition approach is proposed in [10] called a sparse mixture of Gaussian (S-MoG) for debond detection in composites. The algorithm provides reasonable results for the flat shape CFRP. Also, this algorithm is tested for the irregular shape CFRP V-shaped having debond defects at the elbow location.

The problem with existing approaches and algorithm is that as the depth of the defect on the specimen increases or the diameter of the defect on the specimen decreases, the detection performance gets worse and the algorithms fail to detect the defects. Also, when the existing algorithms were tested on an irregular (V-Shape) specimen their performance is not good [10]. The time consumption that is the running time of the algorithms representing their computational efficiency is another problem that needs to be dealt with for meeting the requirements of the online NDT.

In this paper, we solve the task of debonding detection in CFRP composites using optical thermography by using the concatenated feature space where the sparse data, low-rank data, and reconstructed data are used in a joint concatenated matrix. Further, the eigendecomposition is used to represent this joint feature space. It helps embed the sparse and low-rank data in single feature space for optimization. Further, this is solved by an improved tensor nuclear norm (I-TNN) based core singular matrix utilization tensor decomposition framework. This utilization allows exploiting the tensor data in its third mode which is not fully utilized [11] and helps decrease the computational cost of the overall algorithm due to faster convergence. The I-TNN with core matrix decomposition helps extract the weak debond defect information. The proposed approach is compared with general and state-of-the-art optical pulse thermography-based NDT algorithms. The experiment is carried out with two different

CFRP specimens with a flat and irregular shape having multiple debond defects with multiple depths and diameters. The results reflect the efficacy of the proposed algorithm in detecting the weak and noisy defect information from the irregular shape CFRP specimen with less computation time where other algorithms fail.

The rest of the paper is organized as; Section 2 presents the related work; Section 3 presents the proposed algorithm. Section 4 gives information about the experiment setup and specimen. Section 5 presents results and discussion. Section 6 concludes the paper.

II. RELATED WORK

The algorithms used for the debond detection in CFRP using optical thermography can be classified as post-processing techniques that are used to enhance the defect contrast and resolution and to remove the unwanted background noise from the thermal sequences. The debonds are the subsurface defects with varying depths and diameters that are difficult to detect by using only optical thermography due to uneven heating and thermal noise present in the thermal video sequences. To remove this thermal noise such that the defects are visible in the thermal sequences, post-image processing techniques are used. However, by using these techniques the debond defects with small depth and larger diameters are easily detected, but the debond defects with higher depth and smaller diameter are still a challenge. Further, if these debonds are on irregular shape specimens such as elbows and joints, the debond detection problem becomes much more severe. Some of the recent methods and techniques to cater to this problem are discussed below.

In [15], the authors proposed a new excitation method based on the laser thermal excitation for the debonding defect detection in the concrete specimen reinforced by fiber plastics. The authors validate their approach by providing numerical results and feasibility studies. In [16], the authors proposed a deep learning-based thermal image segmentation approach to quantify the debond defects in CFRP using optical thermography. The authors present a temporal and spatial deep network by integrating the cross-network learning strategy. The probability of detection is carried out as a quantitative measure in comparison with other algorithms and experiment results are presented for different CFRPs with debond defects. In [17], the authors present the wavelet feature-based thermal image segmentation for detecting debond defects in the CFRP using optical thermography. The PCA-based features are extracted from the thermal images which are further processed using the Gaussian Low Pass filtering in the wavelet domain. The F-score based comparison is presented with other recent algorithms. In [18], the authors propose a sparse low-rank matrix decomposition method for debond detection in CFRP. A joint decomposition is proposed with iterative sparse modeling. The visual results along with F-score based results are presented to prove the efficacy of the proposed model. In [19], the authors present a comparative study of extracting subsurface defects in thermal patterns. The non-negative matrix factorization methods are used and compared on the thermal data and results are presented in terms of detection accuracies. In [20], the authors present a comparison of tensor-

based defect detection using the eddy current thermography. The importance of using tensor-based algorithms is highlighted. Further matrix and tensor decomposition-based algorithms are also compared. In [21], the authors propose a method of defect depth estimation in a CFRP specimen with flat bottom holes using pulsed thermography. The analysis is presented to characterize the defect depth with regard to specimen thickness and defect size. In [22], the authors propose an automated defect detection method in thermal sequences using important frame selection, feature extraction, and image segmentation to detect the defect size. In [23], the authors propose a Levenberg-Marquardt algorithm to remove the uneven heating noise in thermal sequences. A comparison is also presented with the existing algorithms in terms of noise removal and image resolution. In [24], the authors propose an image segmentation algorithm using artificial intelligence and fuzzy clustering for defect detection in thermal sequences. Experimental analysis is carried out to show the efficacy of the proposed model.

III. PROPOSED METHODOLOGY

First, given the thermographic video sequences $\mathcal{X} \in \mathfrak{R}^{n_1 \times n_2 \times n_3}$, where (n_1, n_2) are the spatial resolution and n_3 is the frame number. In tensor-based terminology, this is a three-way tensor [11], [12]. We propose a multilayer joint decomposition structure [9], [10] of low-rank (L) and sparse components (C) as;

$$\mathcal{X}^1 = \mathcal{L}^1 + \mathcal{C}^1 \quad (1)$$

For the second layer;

$$\mathcal{X}^2 = \mathcal{L}^2 + \mathcal{C}^2 + f^1(\mathcal{X}^1) \quad (2)$$

For the n^{th} layer;

$$\mathcal{X}^n = \mathcal{L}^n + \mathcal{C}^n + f^{n-1}(\mathcal{X}^{n-1}) \quad (3)$$

where $f^n(\mathcal{X}^n)$ is the activation used for the multilayer data modeling. To extract the defect information we propose the following optimization problem [11], [12];

$$\min_{\mathcal{L}, \mathcal{C}} \|\mathcal{L}^n\|_* + \partial \|\mathcal{C}^n\|_1 \text{ s.t } \mathcal{X}^n = \mathcal{L}^n + \mathcal{C}^n \quad (4)$$

where $\|\cdot\|_*$ represents the tensor nuclear norm, ∂ is the regularizing parameter and $\|\cdot\|_1$ is the l_1 norm. The problem in (4) is solved in the concatenated feature space using two steps. In the first step, the low-rank term is solved in the other step the sparse term is solved. Given the thermographic sequences and initializations a concatenated Eigen matrix decomposition is proposed as;

$$\mathbf{Y}^n = \begin{bmatrix} \mathbf{X}^n \\ \mathbf{X}^n - \mathbf{C}^{n-1} \\ \mathbf{C}^{n-1} \end{bmatrix} \quad (5)$$

Where n is the layer number $\mathbf{X} = \beta(\mathcal{X})$ is the tensor to matrix transformation and $\mathcal{X} = \beta^{-1}(\mathbf{X})$ is matrix to tensor transformation, same is used for \mathbf{C} and \mathcal{C} . By joint concatenation of the low-rank and the sparse data, two benefits are obtained. First, we retain the original features of the thermographic sequences which helps prevent the estimated low-rank features to deviate from the original. Further, the residual data and sparse data allow us to embed the sparse

component into low-rank modeling which significantly helps to extract the weak defect information. It should be noted here, that this concatenation firstly occurs in the matrix feature space and the low-rank and sparse modeling is solved in the tensor feature space. This tensor-matrix sparse low-rank decomposition enables us to utilize both tools simultaneously. The problem in (5) is solved using a simple eigendecomposition as;

$$Y^n = UV^T \quad (6)$$

Where U and V are the left and right Eigen matrices and Γ is the diagonal matrix containing the eigenvalues. The first six principal components are selected after the decomposition. Based on the repeated experimental analysis the first six components contain the most useful low-rank information namely.

Here $svt_{\partial_1}(\bar{S})$ is the singular value thresholding for the matrix, $sth(\mathcal{Y})$ is the soft thresholding for tensor [11] and $t - svt_{\frac{\rho}{2}}(\bar{L})$ represents the tensor singular value thresholding for the tensor [12]. For the parameters we set $\partial_1 = \frac{1}{\sqrt{n_{max}}}$, $n_{max} = \max(\min(n_1, n_2), n_3)$. Here (n_1, n_2, n_3) are the frames and rows and columns of core tensor respectively. $\partial = \frac{1}{\sqrt{\max(n_1, n_2)n_3}}$, $\rho = 1.1$ and the stopping condition is set as; $\|\mathcal{L}_{k+1} - \mathcal{L}_k\| \leq \varepsilon$, $\varepsilon = 1e - 5$ based on our experimental analysis. For other applications, the parameters can be tuned accordingly.

$$y^n = \beta^{-1}((UV^T)_{1 \text{ to } 6}) \quad (7)$$

A. Improved TNN with Core Tensor Decomposition

Let $\mathcal{A} = \mathcal{U} * \mathcal{S} * \mathcal{V}^*$ be the tensor singular decomposition (t-svd) of $\mathcal{A} \in \mathfrak{R}^{n_1 \times n_2 \times n_3}$. The tensor nuclear norm of \mathcal{A} is given as [12];

$$\|\mathcal{A}\|_* := \langle \mathcal{S}, \mathcal{J} \rangle = \sum_{i=1}^r \mathcal{S}(i, i, 1) \quad (8)$$

Where $r = rank_t(\mathcal{A})$ is the tensor tubal rank, \mathcal{S}, \mathcal{J} is the singular value and identity tensor. To utilize the information in the third mode of the tensor svd we use the core singular matrix decomposition. The related operators are $\bar{\mathcal{S}} = \beta(\mathcal{S})$ and $\mathcal{S} = \beta^{-1}(\bar{\mathcal{S}})$, where β and β^{-1} are the transforms between the core tensor $\mathcal{S} \in \mathfrak{R}^{n_1 \times n_2 \times n_3}$ and core matrix $\bar{\mathcal{S}} \in \mathfrak{R}^{n \times n_3}$, where $n = \min(n_1, n_2)$.

Further, the most significant components of this core tensor are found by approximating very few significant singular values. The core singular matrix approximation of the I-TNN rank is given as;

$$rank_i(\mathcal{A}) = [rank(\bar{\mathcal{S}}), rank_t(\mathcal{A})]^T \quad (9)$$

Based on the above rank the I-TNN with core matrix approximation is given as;

$$\|\mathcal{A}\|_* = \|\bar{\mathcal{S}}\|_* + \gamma \sum_{i=1}^r \mathcal{S}(i, i, 1) \quad (10)$$

Where γ is the balancing term. Based on the improved tensor nuclear norm and core matrix decomposition the problem in (4) can be solved iteratively using ADMM [13] for each layer.

$$\mathcal{L}_{k+1} = \arg \min_{\mathcal{L}} \|\mathcal{L}\|_* + \frac{\rho}{2} \left\| \mathcal{L} - \mathcal{Y} + \mathcal{C}_k - \frac{\mathcal{D}_k}{\rho} \right\|_F^2 \quad (11)$$

$$\mathcal{C}_{k+1} = \arg \min_{\mathcal{C}} \partial \|\mathcal{C}\|_1 + \frac{\rho}{2} \left\| \mathcal{L}_{k+1} - \mathcal{Y} + \mathcal{C} - \frac{\mathcal{D}_k}{\rho} \right\|_F^2 \quad (12)$$

$$\mathcal{D}_{k+1} = \mathcal{D}_k + \rho(\mathcal{Y} - \mathcal{L}_{k+1} - \mathcal{C}_{k+1}) \quad (13)$$

Where $\rho > 0$ is the augmented lagrangian penalty parameter, \mathcal{D} is the dual variable, k is the iteration number. For the problem in (11), it is solved using ADMM [13] in two steps. The first step solves the core matrix problem and the other.

step solves the tensor nuclear norm problem. The optimization model for the two-step problem can be formulated as [11].

$$\bar{\mathcal{L}}_{k+1} = \arg \min_{\beta(\mathcal{S})} \|\beta(\mathcal{S})\|_* + \frac{1}{2\partial_1} \left\| \mathcal{T} - \mathcal{Y} + \mathcal{C}_k - \frac{\mathcal{D}_k}{\rho} \right\|_F^2 \quad (14)$$

Where δ_1 is the regularizing parameter and \mathcal{S} is from the t-svd and \mathcal{T} is a temporary variable. The tensor with a low-rank core matrix can be given as;

$$\mathcal{Z}_{k+1} = \mathcal{U} * \beta^{-1}(\bar{\mathcal{L}}_{k+1}) * \mathcal{V}^T \quad (15)$$

The other step minimized the tensor nuclear norm as follows;

$$\mathcal{L}_{k+1} = \arg \min_{\mathcal{L}} \|\mathcal{L}\|_* + \frac{\rho}{4} \left\| \mathcal{L} - \mathcal{Z}_{k+1} \right\|_F^2 \quad (16)$$

The details are given in the Table I.

TABLE I. MULTILAYER ADMM FOR I-TNN WITH CORE MATRIX DECOMPOSITION

Input: Tensor data $\mathcal{X} \in \mathfrak{R}^{n_1 \times n_2 \times n_3}$	
Initialization: Given $\rho, \partial_1, \partial, \mathcal{L} = 0, \mathcal{C} = 0, \mathcal{D} = 0$	
For each layer n solve;	
1.	Solve the concatenated problem in (5) by (7) to get \mathcal{Y}
2.	While not converged do:
3.	Compute $[\mathcal{U}, \mathcal{S}, \mathcal{V}] = t - svd(\mathcal{Y} - \mathcal{C} + \frac{\mathcal{D}}{\rho})$
4.	Update $\bar{\mathcal{S}} := \beta(\mathcal{S})$
5.	Compute $\bar{\mathcal{L}} := svt_{\partial_1}(\bar{\mathcal{S}})$
6.	Update $\mathcal{Z} := \beta^{-1}(\bar{\mathcal{L}})$
7.	Update $\bar{\mathcal{L}} := \mathcal{U} * \mathcal{Z} * \mathcal{V}^T$
8.	Update $\mathcal{L} := t - svt_{\frac{\rho}{2}}(\bar{\mathcal{L}})$
9.	Compute $\mathcal{C} := sth_{\frac{\rho}{2}}(\mathcal{Y} - \mathcal{L} + \frac{\mathcal{D}}{\rho})$
10.	Update $\mathcal{D} := \mathcal{D} + \rho(\mathcal{Y} - \mathcal{L} - \mathcal{C})$
11.	End while
Output: $\mathcal{L}^n, \mathcal{C}^n, \mathcal{D}^n$	

IV. EXPERIMENTAL SETUP AND SPECIMEN DETAILS

Regarding the experimental setup, the OPT system available at our lab can be seen in Fig. 1. In the OPT system, halogen lamps are used as an excitation source to induce heat into the specimen. A $ZY - B$ type excitation source is used at the back of the halogen lamps with model *ITECH - IT6726G*. This model supports an adjustable power source that can be

tuned up to 3kW. In our experiments, we set the power of the excitation source to 2kW. The experiments are conducted in the reflection mode configuration [14]. The 85cm distance is selected between the lamps and the specimen. To capture the thermographs of the sample an infrared camera is used model A655sc. The output resolution of the infrared camera is set to 640 × 480 per frame. The infrared camera used has a thermal sensitivity of 0.05°C. Further, a sampling frequency of 50Hz is selected for our experiments.

To validate the proposed model, two different CFRP specimen is selected. The CFRP specimen was obtained from the Chengdu Aircraft Design Institute of the China Aviation Industry. The obtained specimen is similar to the ones used in the design and manufacturing of aircraft and related components. The first specimen has a rectangular shape and a flat surface. The second sample is like a V shape with joints and edges. Both the specimen have debonded defects created

by inducing the Teflon inserts at various depths and different diameters. The detailed information about the specimen shape and defects can be found in Table II.

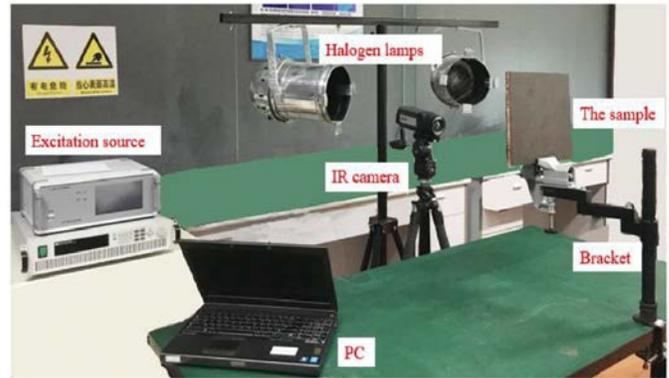


Fig. 1. The Optical Pulse Thermography System.

TABLE II. INFORMATION ABOUT THE CFRP SPECIMEN

Number	Defect Profile	Dimension(mm)	Defect Information(mm)	Picture
1		250×250×24.2	Depth: 1, 2 Diameter: 2,4,6,8,10,12,16, 20	
3		100×100×80	Depth: 1.5, 1.75, 2, 2.25, 2.5, 2.75 Diameter: 9, 10	

V. RESULTS AND DISCUSSION

In this section, the results are presented from the experiments. The visual results along with the F-score [9] and computation time are presented. The experiments are carried out using the optical pulse thermography system shown in Fig. 1. The repeated experiments were carried out to collect the thermographic sequences used. The reflection mode configuration is used with the standard camera, excitation source, and specimen distance given in [9]. The algorithms used in competition with the proposed algorithm are PPT [4], TSR [3], SPCT [6], EVBTF [9], and S-MoG [10]. The F-score and computation time results are given in Table III. The proposed algorithm was run up to four layers and the third layer results were found to be clear with better resolution and contrast for detecting the debond defects.

The defect detection results from the algorithms in comparison are given in a tabular form shown in Fig. 2. All the results for both specimens can be seen in Fig. 2. Fig. 2(Row 1) shows the comparative results for specimen 1. It is a rectangular-shaped specimen having a flat surface and 10 debond defects. The depths of the debonds are (1mm and 2mm) with varying diameter sizes. All the algorithms are run on a single computer to avoid any unfair processing advantage. From Fig. 2 (Row 1) it can be seen that most of the algorithms can detect at most 9 out of 10 defects. However, the strong noise is still present which limits the performance of algorithms in the scenario when the information about the defects is unknown. The proposed algorithm gives reasonable results with a reduction in noise and improvement in the resolution contrast of the specimen. The proposed algorithm can detect all defects where other algorithms fail.

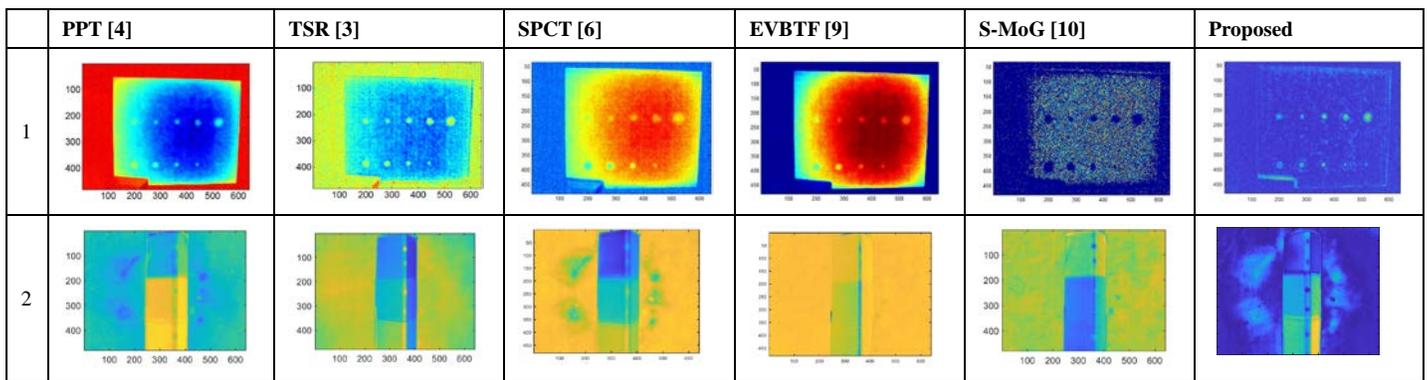


Fig. 2. The Comparative Analysis of different Algorithms.

TABLE III. COMPARATIVE RESULTS F-SCORE (LEFT) AND TIME TAKEN (RIGHT IN SECONDS)

Specimen Number	PPT [4]		TSR [3]		SPCT [6]		EVBTF [9]		S-MoG [10]		Proposed	
1	0.94	135	0.94	271	0.94	40	0.94	1342	0.94	173	1	26
2	0.75	146	0.88	601	0.88	43	0.75	753	0.88	125	1	6
Average	0.84	140	0.91	436	0.91	41	0.84	1046	0.91	149	1	16

Fig. 2(Row 2) shows the results for the second specimen. It has a V shape with an irregular surface. It is a more challenging specimen with defect depths ranging from (1.5, 1.75, 2, 2.25, 2.5, 2.75)mm. Here, the defect diameters are (9 and 10)mm. It can be seen from Fig. 2(Row 2) that almost all algorithms fail to detect the debond defects. The strong noise is present and the resolution of the defects is quite poor. The proposed algorithm can detect all the defects with reasonable noise reduction and acceptable resolution. As the proposed algorithm utilizes a multilayer structure, the results in Fig. 2 are obtained by two layers of the algorithm. This layer number is selected based on the experimental analysis with multiple experiments on different data. However, due to the concatenated feature space where low-rank, sparse, and raw data are decomposed using a single feature space further layering does not significantly improve the results but in turn, incurs the additional computation cost. From the visual results, it can be argued that the proposed I-TNN with core matrix decomposition using a concatenated feature space in a multilayer architecture can detect smaller and deeper debond defects using OPT-NDT.

Table III shows the F-score and computation time results. For both specimens, the results are averaged in the last row of the Table III. On average, the PPT [4] and EVBTF [9] algorithms have the detection efficiency of 84% with 140sec and 1046sec as the average computation time. The TSR [3], SPCT [6] and S-MoG [10] algorithms have the 91 average percent of defect detection. Their running times are 436sec, 41 sec and 149sec respectively. The proposed model has 100% defect detection accuracy for the specimen under test. Using the concatenated feature space a simple eigen decomposition is carried out and only 6 principal eigen components are selected which help significantly in reducing the computation time enhancing the resolution of defects by multilayer I-TNN core matrix decomposition approach. The average computation time taken by the proposed model is 16sec.

VI. CONCLUSION AND FUTURE WORK

In this paper, a joint concatenation of the low-rank and sparse features is proposed. The I-TNN algorithm with core matrix-tensor decomposition in multilayer architecture is proposed for iteratively solving the feature space. The proposed approach is used for detecting the debond defects in the CFRP specimen using optical pulse thermography. By multi-layering, the low-rank and sparse components in a concatenated feature space help boost the convergence, eliminating the noise and detecting the debond defects with small diameter and varying depth. The comparative analysis with general OPT-NDT and other low-rank sparse and tensor modeling algorithms proves the debond detection capability of the proposed algorithm.

The possible future extensions of this work will be the testing of this work on several new and different CFRP specimens. Further, the proposed algorithm can be tested and validated on the eddy current thermography data, microwave thermography data, and other thermography data. Further apart from the debond detection the algorithm can also be used for other defects using as delaminations and crack in CFRP and other metal structures.

REFERENCES

- [1] S. Marinetti et al., "Statistical analysis of IR thermographic sequences by PCA," *Infrared Phys. Technol.*, vol. 46, no. 1–2, pp. 85–91, 2004.
- [2] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4–5, pp. 411–430, 2000.
- [3] S. M. Shepard and M. F. Beemer, "Advances in thermographic signal reconstruction," in *Thermosense: Thermal Infrared Applications XXXVII*, 2015, vol. 9485, pp. 1–7.
- [4] X. Maldague and S. Marinetti, "Pulse phase infrared thermography," *J. Appl. Phys.*, vol. 79, no. 5, pp. 2694–2698, 1996.
- [5] C. Ibarra-Castanedo and X. Maldague, "Pulsed phase thermography reviewed," *Quant. Infrared Thermogr. J.*, vol. 1, no. 1, pp. 47–70, 2004.
- [6] B. Yousefi, S. Sfarra, F. Sarasini, and X. P. V Maldague, "IRNDT inspection via sparse principal component thermography," in *2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE)*, 2018, pp. 1–4.

- [7] J.-Y. Wu, S. Sfarra, and Y. Yao, "Sparse principal component thermography for subsurface defect detection in composite products," *IEEE Trans. Ind. Informatics*, vol. 14, no. 12, pp. 5594–5600, 2018.
- [8] L. Zhang, Q. Zhang, L. Zhang, D. Tao, X. Huang, and B. Du, "Ensemble manifold regularized sparse low-rank approximation for multiview feature embedding," *Pattern Recognit.*, vol. 48, no. 10, pp. 3102–3112, 2015.
- [9] P. Lu, B. Gao, Q. Feng, Y. Yang, W. L. Woo, and G. Y. Tian, "Ensemble variational Bayes tensor factorization for super resolution of CFRP debond detection," *Infrared Phys. Technol.*, vol. 85, pp. 335–346, 2017.
- [10] J. Ahmed, B. Gao, G. Y. Tian, Y. Yang, and Y. C. Fan, "Sparse ensemble matrix factorization for debond detection in CFRP composites using optical thermography," *Infrared Phys. Technol.*, vol. 92, pp. 392–401, 2018.
- [11] Y. Liu, L. Chen, and C. Zhu, "Improved Robust Tensor Principal Component Analysis via Low-Rank Core Matrix," *IEEE J. Sel. Top. Signal Process.*, vol. 12, no. 6, pp. 1378–1389, 2018.
- [12] C. Lu, J. Feng, W. Liu, Z. Lin, and S. Yan, "Tensor robust principal component analysis with a new tensor nuclear norm," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.
- [13] R. Liu, Z. Lin, and Z. Su, "Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning," in *Asian Conference on Machine Learning*, 2013, pp. 116–132.
- [14] Y. He, R. Yang, H. Zhang, D. Zhou, and G. Wang, "Volume or inside heating thermography using electromagnetic excitation for advanced composite materials," *Int. J. Therm. Sci.*, vol. 111, pp. 41–49, 2017.
- [15] Y. Xu and H. Sohn, "Nondestructive debonding detection of fiber reinforced plastics strengthened structure based on infrared thermal imaging with laser thermal excitation," in *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2020*, 2020, vol. 11379, p. 1137914.
- [16] Q. Luo, B. Gao, W. L. Woo, and Y. Yang, "Temporal and spatial deep learning network for infrared thermal defect detection," *NDT E Int.*, vol. 108, p. 102164, 2019.
- [17] J. Ahmed, G. A. Baloch, and G. Y. Tian, "Wavelet Domain Based Defect Detection using Optical Thermography," in *ACM International Conference Proceeding Series*, 2019, pp. 1–5, doi: 10.1145/3332340.3332356.
- [18] J. Ahmed, B. Gao, W. L. Woo, and Y. Zhu, "Ensemble Joint Sparse Low-Rank Matrix Decomposition for Thermography Diagnosis System," *IEEE Trans. Ind. Electron.*, vol. 68, no. 3, pp. 2648–2658, 2020.
- [19] B. Yousefi, C. I. Castanedo, and X. P. V. Maldague, "Measuring heterogeneous thermal patterns in infrared-based diagnostic systems using sparse low-rank matrix approximation: Comparative study," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2020.
- [20] Y. Liang, L. Bai, J. Shao, and Y. Cheng, "Application of Tensor Decomposition Methods In Eddy Current Pulsed Thermography Sequences Processing," in *2020 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, 2020, pp. 401–406.
- [21] Y. Wei, S. Zhang, Y. Luo, L. Ding, and D. Zhang, "Accurate depth determination of defects in composite materials using pulsed thermography," *Compos. Struct.*, vol. 267, p. 113846, 2021.
- [22] L. Yuan, X. Zhu, Q. Sun, H. Liu, P. Yuen, and Y. Liu, "Automatic extraction of material defect size by infrared image sequence," *Appl. Sci.*, vol. 10, no. 22, p. 8248, 2020.
- [23] D. Wang, Z. Wang, J. Zhu, and F. Ciampa, "Enhanced pre-processing of thermal data in long pulse thermography using the Levenberg-Marquardt algorithm," *Infrared Phys. Technol.*, vol. 99, pp. 158–166, 2019.
- [24] Z. Wang, L. Wan, N. Xiong, J. Zhu, and F. Ciampa, "Variational level set and fuzzy clustering for enhanced thermal image segmentation and damage assessment," *NDT E Int.*, vol. 118, p. 102396, 2021.

Unmoderated Remote Usability Testing: An Approach during Covid-19 Pandemic

Ambar Relawati¹, Yanuar Primanda³

School of Nursing, Universitas Muhammadiyah Yogyakarta
Yogyakarta, Indonesia

Guntur Maulana Zamroni²

Department of Informatics Engineering
Universitas Ahmad Dahlan, Yogyakarta, Indonesia

Abstract—Online Nurse Test for Indonesian Nurse Competency (ONT UKNI) is a mobile application that was developed to help increase the success rate of nurse competency test participants. By using this application, users can learn more about the materials tested and conduct try out as a competency test simulation. However, ONT UKNI has not yet passed adequate testing stages, especially in terms of User Interface/User Experience (UI/UX). The Covid-19 pandemic situation presents challenges in the UI/UX testing process. Testing process which is ideally carried out face-to-face with respondents to get further insight, have to be carried out using another approach following the new normal protocol. This study aims to test the usability of UI/UX with an unmoderated remote testing approach on ONT UKNI application using a USE questionnaire. The test was performed using 26 respondents and all were nursing profession students of Universitas Muhammadiyah Yogyakarta. Respondents performed 8 tasks on ONT UKNI and answered set of questionnaire that will be tabulated and analyzed. The results indicate that usefulness, ease of learning, and satisfaction variables get the Very Good category while the ease of use variable gets the Good category. Overall, usability testing using an unmoderated remote testing approach can be carried out and able to provide information about areas where users are satisfied with ONT UKNI application. However, some areas still have room for improvement such as better UI design and implementation of gamification.

Keywords—Mobile learning; nurse competency test; unmoderated remote testing; usability

I. INTRODUCTION

Mobile-based learning media is not a new thing in an effort to improve the quality of educational outcomes. Rapid growth of mobile technology and its application gradually replace the role of computer [1]. The author in [2] conducted research on the effectiveness of using Android-based learning media for biology subjects in high school students. From the research conducted, it is known that the learning media has a positive impact. Android-based learning media increases interest and motivation in learning which indirectly increases the effectiveness of learning outcomes. Android-based learning media provides several advantages such as attractive designs, both in terms of images, colors, and writing. Learning media is also easy to use independently either at school or outside of school. In another study conducted by [3], it is known that school students are easier to accept learning materials in digital form than in written or oral form. Digital learning media such as Android-based learning media application have several advantages, such as ease of use, can be used for learning

anywhere, and can be used offline [4]. This shows that mobile applications are increasingly popular for use in education area.

Online Nurse Test for Indonesian Nurse Competency (ONT UKNI) is a mobile application that was developed and aimed to increasing the success rate of nurse competency test participants [5]. By using this application, users –in this case are nursing profession students- can learn about the materials tested at competency test and conduct try out as competency test simulations. However, ONT UKNI has not passed adequate testing stages, especially in terms of user experience or User Interface / User Experience (UI/UX). This is important because user experience determines the success or failure of a product [6]. If product's usability cannot satisfy user, it will hinder the overall quality of the application. Thus developer should consider usability when design an application so it will meet its purpose [7]. Furthermore, [8] added that the mobile app needs to go through a thorough testing phase due to several factors such as the limited screen size. For this reason, it is necessary to conduct user-oriented testing to ensure feasibility and user experience.

Usability testing can be used to test feasibility and user experience [9]. The author in [10] conducted research related to usability testing on the online guardianship application STMIK AMIK Bandung by using a USE questionnaire as a research instrument. Although USE questionnaire lack of evidence in reliability and validity, it provides insight for researchers in terms of usefulness, satisfaction, and ease of use [11][12]. The author in [13] conducted usability testing research on YouTube websites among Malaysian teenagers using the same 3 criteria, namely usefulness, satisfaction, and ease of use. Apart from the USE questionnaire, there are also several other usability testing instruments, such as the User Experience Questionnaire (UEQ) and the System Usability Scale (SUS) [14][15]. The author in [16] performed usability testing on the MyTelkomsel mobile application using 5 criteria, namely: learnability, efficiency, memorability, error, and satisfaction. The data collection technique used was a survey with SUS instrument, observation, and direct interviews. Based on this research, it can be seen that the interview technique helped researchers to get further insight and find solutions to the problems that respondents complains about the MyTelkomsel application. The author in [17] also used a survey and interview techniques in usability testing with certain variables, such as demographic information, using experience, ease of use, and usefulness.

The Covid-19 pandemic situation presents challenges in the UI/UX testing process. The author in [18] even stated that

based on a study conducted by researchers from Harvard University, Covid-19 protocols such as physical distancing will last until 2022. This conditions certainly become obstacles and challenges in various sectors, including the software testing sector. UI/UX testing, which ideally be carried out in-person with respondents to get optimum feedback, should be carried out using another approach following the new normal protocol policy [19][20]. This can be overcome by conducting UI/UX testing with an unmoderated remote usability testing approach [21]. Unmoderated remote usability testing offers flexibility within time and distance constraints.

Based on the explanation above, the researcher decided to conduct usability testing on ONT UKNI application using an unmoderated remote testing approach. Unmoderated testing is a test where respondents try the product being tested and provide an assessment without being accompanied by a moderator [22]. The moderator creates a test scenario for respondents to follow who will then provide an assessment. Remote testing approach can solve problems related to distance since respondents can be located anywhere [23][24]. Thus, remote testing considered suitable to be used considering physical distancing protocol during Covid-19 pandemic situation. USE questionnaire instrument was used to obtain qualitative data then processed further to determine user experience in terms of usefulness, satisfaction, ease of use, and ease of learning. This research contributes to theoretical and practical basis of using USE questionnaire in unmoderated remote testing and encourages professionals to adopt this approach.

II. ONT UKNI BACKGROUND

Nursing education is one area that can benefited of Android applications as a learning media for the preparation of the Indonesian Nurses Competency Test (UKNI). UKNI is a process to determine whether or not someone is eligible to become a nurse in Indonesia [25]. The author in [26] stated that the graduation rate in 2018 was 57.1% with 26,208 graduates, in 2019 it was 58.6% with 29,240 graduates, and in 2020 it was 54.4% with 23,663 graduates. This figure is quite low considering the need for the nursing profession in Indonesia is quite high. Indonesian Central Statistics Agency predicts that the need for nurses in Indonesia is 48,253,428 nurses with a ratio of 180 nurses per 100,000 population [27]. The need for an additional number of nurses in Indonesia in 2019 is predicted to be 372,534 nurses [28]. This is in line with the statement of the Professor of the Faculty of Nursing Universitas Indonesia, Achir Yani Hamid, who stated that the number of Indonesian nursing personnel was only 60% of the total population where the world standard was 80% [29].

The author in [30] stated that the obstacles faced by UKNI participants were lack of focus and lack of time to prepare, considering that most UKNI participants were students who were practicing the nursing profession. Another reason was the participants' ignorance about the UKNI concept, especially regarding the UKNI test grid. The author in [31], [32], and [33] analyze the factors that influence UKNI graduation, including try out, Grade Point Average, learning style, and learning motivation.

ONT UKNI was developed to overcome problems that have been stated before. It is a mobile application which has features such as learning material and try out simulation as shown in Fig. 1. Learning materials and try out given have been adapted to the material being tested at UKNI to provide real experience regarding nurse competency test.

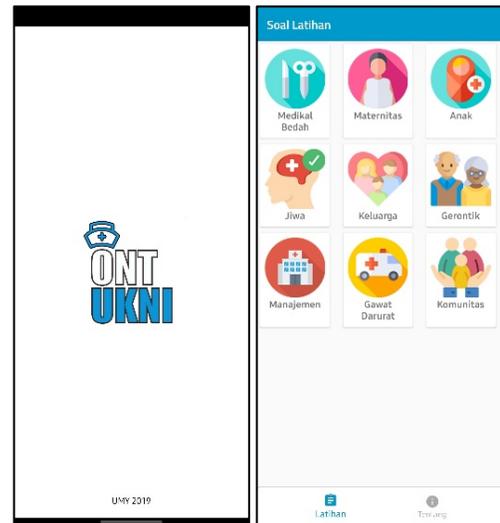


Fig. 1. ONT UKNI Mobile Application.

III. METHODOLOGY

This study used a four-step procedure. Fig. 2 shows the research methodology used in this study. The authors start from literature review, requirement analysis, usability testing and data tabulation, and analysis.

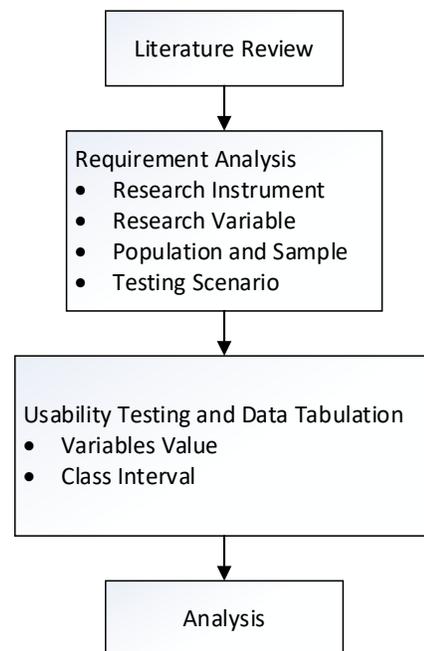


Fig. 2. Research Methodology.

A. Literature Review

The authors examine several theories sourced from books, reputable online sources, or journals. Several theories used in

this research include usability testing, USE questionnaire, testing, population, sample, data tabulation. This stage also aims to explore and find references related to previous studies.

B. Requirement Analysis

Research instruments and variables will be determined in this phase. The research instrument used to perform usability testing on a number of respondents. Sample size of respondents were taken from the population who are potential users of ONT UKNI application. Respondents will be asked to perform several steps contained in the test scenario that has been given by the author. Respondents were given 30 days to try the ONT UKNI application with the aim of getting as much user experience as possible. After that the respondents will fill out a questionnaire provided by the author.

C. Usability Test and Data Tabulation

At this stage the respondents will fill out an online assessment questionnaire based on the experience they have gained while using ONT UKNI application. The usability testing process is carried out remotely using Google Forms and is not moderated by researchers. Calculation of the value of each variable used in usability testing is carried out. Each variable used will be categorized based on class intervals.

D. Analysis

Usability test and data tabulation results were analyzed to find out how the results of each research variable were used. Recommendations regarding applications were analyzed to give ONT UKNI developer insight.

IV. RESULTS AND DISCUSSION

A. Requirement Analysis

1) *Research instruments and variables:* At this stage the authors determine the instruments and research variables used. The assessment scale used in the questionnaire is Likert scale with descriptive statements converted into numerical values as in Table I. USE questionnaire developed by Arnold M. Lund in 2001 was used as a test instrument [12]. USE questionnaire consists of 30 statements which are divided into four variables of assessment variables, namely: usefulness, ease of use, ease of learning, and satisfaction. Usability instrument has been modified according to the scope of the research as shown in Tables II to V.

TABLE I. LIKERT SCALE ASSESSMENT CRITERIA

Statement	Mark
Strongly agree	5
Agree	4
Neutral	3
Disagree	2
Strongly Disagree	1

2) *Research population:* The population in this study were prospective users of ONT UKNI application, namely students

of the nursing profession at the Universitas Muhammadiyah Yogyakarta who were at PKU Yogyakarta Hospital. This population was chosen because UKNI is a test for professional students to get nurse certification. Thus respondents were in a preparation for UKNI. From the entire population of nursing profession students of Universitas Muhammadiyah Yogyakarta who were interns at PKU, several samples taken to become respondents in usability testing.

TABLE II. USEFULNESS QUESTIONNAIRE INSTRUMENT

No	Questions
U1	Does ONT UKNI application help me to be more effective in preparing for UKNI?
U2	Does ONT UKNI application help me to be more productive in preparing for UKNI?
U3	ONT UKNI application is very useful in preparing for UKNI.
U4	ONT UKNI application makes me more flexible in managing my study time to face UKNI.
U5	ONT UKNI application makes the learning process to deal with UKNI easier to do.
U6	ONT UKNI application saves me time in preparation for UKNI.
U7	ONT UKNI application fulfills my needs in preparation for UKNI.
U8	ONT UKNI app did a lot of what I expected with regards to UKNI preparation.

TABLE III. EASE OF USE QUESTIONNAIRE INSTRUMENT

No	Questions
EU1	ONT UKNI application is easy to use.
EU2	ONT UKNI application is practical to use.
EU3	ONT UKNI application is user friendly.
EU4	ONT UKNI application requires a few steps when I want to use a feature in the application.
EU5	This application is very flexible.
EU6	Using ONT UKNI application does not require a lot of effort.
EU7	I can use ONT UKNI application without requiring written instructions or guidance.
EU8	I did not find any inconsistencies in ONT UKNI application.
EU9	Users who regularly or only occasionally use ONT UKNI application will love this application.
EU10	I can solve errors in using ONT UKNI application easily.
EU11	I managed to use ONT UKNI app every time.

TABLE IV. EASE OF LEARNING QUESTIONNAIRE INSTRUMENT

No	Questions
EL1	I was able to quickly learn how to use ONT UKNI app.
EL2	I can easily remember how to use ONT UKNI application.
EL3	ONT UKNI application is easy to learn.
EL4	I quickly became proficient in using ONT UKNI application.

TABLE V. SATISFACTION QUESTIONNAIRE INSTRUMENT

No	Questions
S1	I am satisfied with ONT UKNI application.
S2	I will recommend ONT UKNI application to my friends.
S3	Using ONT UKNI application is very fun.
S4	ONT UKNI app worked as I expected.
S5	ONT UKNI application has a nice display.
S6	I feel the need to use ONT UKNI application in preparation for UKNI.
S7	ONT UKNI application is convenient to use in preparing for UKNI.

3) *Sample size:* The sampling technique used is simple random sampling where each member of the population has the same opportunity to become a respondent. The size of sample is determined using the Slovin formula approach as described as in (1).

$$n = \frac{N}{1+N(e)^2} \quad (1)$$

Where:

n = Sample size

N = Population Size

e = error rate

The population size of nursing profession students at Muhammadiyah Yogyakarta University who are at PKU Yogyakarta Hospital is 27 students. The error rate used in the search for the sample size is 5%. By using (1), it can be calculated the sample size used in the usability test. The sample size used was 25.292 respondents which were rounded up to 26 respondents.

$$n = \frac{27}{1+27(0,05)^2} = 25,292$$

4) *Testing scenario:* Testing is carried out without moderation and remotely considering the Covid-19 pandemic situation. Remote testing is possible in the presence of a test scenario [34]. Test scenario consists of 8 steps in a task-based form and aims to ensure that respondents have tried and understood the features of ONT UKNI application. Table VI shows the scenarios that will be carried out by respondents remotely without moderation. Test scenario is made as simple as possible to make it easier for respondents to follow and does not require any clarification from researchers [24][35]. 26 respondents in this study carried out a number of activities

following the scenario given by the research team. Respondents were given 14 days to try ONT UKNI application and carry out the given scenario and try ONT UKNI application. After 14 days, the respondent will assess the usability of ONT UKNI application using the provided instrument.

B. Usability Testing and Data Tabulations

Tables VII to X shows post-usability test results that were tabulated and compiled. From the test results, it can be calculated the value of each variable used. Equation (2) used to determine the usability variable, ease of use, ease of learning, and satisfaction value.

$$\text{Nilai Variabel} = \frac{s}{nr \times np} \quad (2)$$

Where:

s = Total value

nr = Number of respondents

np = Number of statements

$$\text{Usefulness variable value} = \frac{876}{26 \times 8} = 4,21$$

$$\text{Ease of Use variable value} = \frac{1190}{26 \times 11} = 4,16$$

$$\text{Ease of Learning variable value} = \frac{447}{26 \times 4} = 4,29$$

$$\text{Satisfaction variable value} = \frac{766}{26 \times 7} = 4,20$$

TABLE VI. TASK SCENARIO

No	Scenario
1	Downloading the ONT UKNI application apk that has been shared
2	Installing ONT UKNI application
3	Running ONT UKNI application
4	Try the study menu
5	Try the training menu
6	Checking the value of the results of the exercise and discussion
7	Try the menu about developer
8	Exit UKNI ONT app

TABLE VII. USEFULNESS (U)

No	Likert Scale					N	Min	Max	Mean	SD
	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree					
U1	8	16	1	0	1	26	1	5	4,15	0,818
U2	6	18	1	0	1	26	1	5	4,08	0,781
U3	9	15	2	0	0	26	3	5	4,27	0,592
U4	8	16	1	1	0	26	2	5	4,19	0,680
U5	9	14	2	1	0	26	2	5	4,19	0,735
U6	11	14	1	0	0	26	3	5	4,38	0,560
U7	7	17	1	1	0	26	2	5	4,15	0,662
U8	9	16	0	1	0	26	2	5	4,27	0,654

TABLE VIII. EASE OF USE (EU)

No	Likert Scale					N	Min	Max	Mean	SD
	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree					
EU1	10	13	2	1	0	26	1	5	4,23	0,750
EU2	6	18	2	0	0	26	1	5	4,15	0,533
EU3	11	13	1	1	1	26	3	5	4,31	0,722
EU4	7	17	1	1	1	26	2	5	4,15	0,662
EU5	6	18	2	0	0	26	3	5	4,15	0,533
EU6	5	20	0	1	0	26	2	5	4,12	0,577
EU7	7	19	0	0	0	26	4	5	4,27	0,444
EU8	6	17	1	1	1	26	1	5	4,00	0,877
EU9	7	16	3	0	0	26	3	5	4,15	0,601
EU10	6	18	2	0	0	26	3	5	4,15	0,533
EU11	7	15	3	1	0	26	2	5	4,08	0,730

TABLE IX. EASE OF LEARNING (EL)

No	Likert Scale					N	Min	Max	Mean	SD
	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree					
EL1	8	16	1	1	0	26	1	5	4,19	0,680
EL2	11	15	0	0	0	26	4	5	4,42	0,494
EL3	10	14	2	0	0	26	3	5	4,31	0,606
EL4	9	15	2	0	0	26	3	5	4,27	0,592

TABLE X. SATISFACTION (S)

No	Likert Scale					N	Min	Max	Mean	SD
	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree					
S1	7	16	0	2	1	26	1	5	4,00	0,961
S2	8	16	2	0	0	26	3	5	4,23	0,576
S3	8	15	3	0	0	26	3	5	4,19	0,621
S4	8	16	1	0	1	26	1	5	4,15	0,818
S5	7	17	1	1	0	26	2	5	4,15	0,662
S6	11	14	1	0	0	26	3	5	4,38	0,560
S7	10	15	1	0	0	26	3	5	4,35	0,551

To be able to determine the category of assessment class based on known variable values, a class interval size is needed [36]. Equation (3) shows the formula used to calculate the length of the class interval.

$$P = \frac{Dx - Dy}{n} \quad (3)$$

Where:

P = Interval length

Dx = Biggest value

Dy = Smallest value

n = Number of classes

$$P = \frac{5 - 1}{5}$$

$$P = 0,8$$

Based on $P = 0.8$ obtained from (3), an assessment table of the results of the usability test analysis can be made which is then used to determine the category of each variable. Table XI shows the length of the interval for each category. Table XII shows the results of the analysis carried out. The usability variable gets a value of 4.21 and is included in the good category class interval. The ease of use variable gets a value of 4.16 and is included in the good class interval. The ease of learning variable got a value of 4.29 and was included in the good class interval. The satisfaction variable gets a value of 4.20 and is included in the very good class interval.

TABLE XI. CLASS INTERVAL LENGTH

No	Interval Length	Category
1	1.00 – 1.79	Very low
2	1.80 – 2.59	Low
3	2.60 – 3.39	Pretty good
4	3.40 – 4.19	Good
5	4.20 – 5.00	Very good

TABLE XII. VARIABLE CATEGORY BASED ON INTERVAL

No	Variable	Total Value	Average Value	Category
1	Utility	876	4.21	Very good
2	Ease of Use	1190	4.16	Good
3	Ease of Learning	447	4.29	Very good
4	Satisfaction	766	4.20	Very good
Average variable value			4.21	Very good

Usefulness variable gets a value of 4.21 and is included in the Very Good category. ONT UKNI application has learning and tries out features that are in accordance with the needs of application users who are students of the nursing profession in preparation for joining UKNI. In addition, ONT UKNI application has discussion feature so that users will better understand the questions and exact answers given which are very useful, efficient, and can increase user productivity in the preparation for UKNI. One aspect that needs to be considered is the aspect of the responsiveness. Application responsiveness is one of the factors that can further increase user productivity and efficiency in using applications [7].

Ease of use variable gets a value of 4.16 and falls into the Good category. Button design and layout is easy to reach. It has minimalist feature by having 2 main buttons for the main features, namely the learning feature and the try out feature. Users can also use the application even though it is not accompanied by a user guide and at any time. Consistent selection of text and colors also helps users. However, there are still some areas that could be improved. Color selection and interface design are 2 factors that have the potential to be improved to provide better interaction and user experience [37].

Ease of learning variable gets a value of 4.29 and was included in the Very Good category. It shows that ONT UKNI easy for user to understand. It is intuitive enough for first time user to follow. The application has a consistent design which helps user to get familiar faster which is an important aspect in ease of learning [38]. The placement of buttons accompanied by text makes it easier for users in the learning process to use

this application. In addition, the selection of clear and contrasting colors between the buttons, written text, and background also helps the use of ONT UKNI application.

Satisfaction variable gets a value of 4.20 and is included in the Very Good category. ONT UKNI application was built with the aim of helping students of the nursing profession in dealing with UKNI. Based on the test results, it can be seen that users feel confident to use ONT UKNI application in preparation for UKNI. Application design that is attractive, easy to use, and easy to learn is also a factor that determines user satisfaction [39].

V. CONCLUSION AND FUTURE WORK

This study was aimed to conduct usability testing on ONT UKNI application. The test has been carried out on 26 respondents who are nursing profession students of Universitas Muhammadiyah Yogyakarta who were interns at PKU Yogyakarta Hospital. The test was conducted using a USE questionnaire remotely and without any moderation from the research team. USE questionnaire gives insight on 4 variables, namely usefulness, ease of use, ease of learning, and satisfaction. The usability variable gets a value of 4.21 and is included in the Very Good category. The ease of use variable gets a value of 4.16 and falls into the Good category. The ease of learning variable got a value of 4.29 and was included in the Very Good category. The satisfaction variable gets a value of 4.20 and is included in the Very Good category. However, there are several areas to improve the user experience of ONT UKNI application, such as color selection, application interface design, and implementation of gamification into ONT UKNI application. Overall, usability testing with unmoderated remote testing approach using USE questionnaire can be done and able to provide information about areas that users are satisfied with or areas that need improvement.

Future work should focus on improving ONT UKNI application by redesign UI and implementing gamification. Further test with intervention using quasi experiment also possible to be conducted to see whether ONT UKNI application really meets its purpose, which is to help nursing profession student preparing for UKNI.

ACKNOWLEDGMENT

The authors gratefully acknowledge Universitas Ahmad Dahlan for funding assistance in this research and School of Nursing Universitas Muhammadiyah Yogyakarta for research collaboration.

REFERENCES

- [1] M. Elgan, "With Smartphones Like This, Why Do We Need Laptops?," 2017. <https://www.computerworld.com/article/3241233/smartphones-with-smartphones-like-these-why-do-we-need-laptops.html> (accessed Jan. 24, 2019).
- [2] S. Muyaroah and M. Fajartia, "Pengembangan Media Pembelajaran Berbasis Android dengan menggunakan Aplikasi Adobe Flash CS 6 pada Mata Pelajaran Biologi Abstrak," vol. 6, no. 2301, pp. 79–83, 2017.
- [3] A. Teodorescu, "Mobile learning and its impact on business English learning," *Procedia - Soc. Behav. Sci.*, vol. 180, no. November 2014, pp. 1535–1540, 2015, doi: 10.1016/j.sbspro.2015.02.303.

- [4] J. Kuswanto and F. Radiansah, "Media Pembelajaran Berbasis Android Pada Mata Pelajaran Sistem Operasi Jaringan Kelas XI," *An Nabighoh J. Pendidik. dan Pembelajaran Bhs. Arab*, vol. 14, no. 01, p. 129, 2018.
- [5] A. Relawati and G. M. Zamroni, "Development of Android Based Online Nurse Test Preparation," *JUITA J. Inform.*, vol. 8, no. 1, p. 111, 2020, doi: 10.30595/juita.v8i1.6795.
- [6] B. Iglar, T. Braumann, and S. Bohm, "Evaluating the Usability of Mobile Applications Without Affecting The User And The Usage Context Bodo," *Int. J. Bus. Manag. Stud.*, vol. 5, no. 1, pp. 92–102, 2013.
- [7] N. D. Lynn, A. I. Sourav, and D. B. Setyohadi, "Increasing user satisfaction of mobile commerce using usability," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 8, pp. 300–308, 2020, doi: 10.14569/IJACSA.2020.0110839.
- [8] S. A. Wicaksono, D. R. Firdausy, and M. C. Saputra, "Usability Testing on Android Application of Infrastructure and Facility Reporting Management Information System," *J. Inf. Technol. Comput. Sci.*, vol. 3, no. 2, pp. 184–193, 2018, doi: 10.25126/jitecs.20183267.
- [9] W. A. Kusuma, V. Noviasari, and G. I. Marthasari, "Analisis Usability dalam User Experience pada Sistem KRS- Online UMM menggunakan USE Questionnaire," vol. 5, no. 4, pp. 294–301, 2016.
- [10] K. Aelani and K. Kunci, "Pengukuran Usability Sistem Menggunakan Use Questionnaire (Studi Kasus Aplikasi Perwalian Online STMIK " AMIKBANDUNG ")," *Semin. Nas. Apl. Teknol. Inf.*, vol. 2012, no. Snati, pp. 1907–5022, 2012, [Online]. Available: <http://journal.uii.ac.id/Snati/article/viewFile/2913/2676>.
- [11] M. Gao, P. Kortum, and F. Oswald, "Psychometric evaluation of the USE (usefulness, satisfaction, and ease of use) questionnaire for reliability and validity," *Proc. Hum. Factors Ergon. Soc.*, vol. 3, pp. 1414–1418, 2018, doi: 10.1177/1541931218621322.
- [12] A. M. Lund, "Measuring usability with the USE questionnaire," *Usability interface*, vol. 8, no. 2, pp. 3–6, 2001, doi: 10.1177/1078087402250360.
- [13] M. Nur, F. Abd, A. Hussain, M. Maizan, and F. Hamdi, "Usability study of youtube websites for Malaysian teenagers Usability Study of YouTube Websites For Malaysian Teenagers," vol. 020121, no. October, 2017.
- [14] T. Yuliyana, I. K. R. Arthana, and K. Agustini, "Usability Testing pada Aplikasi POTWIS," *JST (Jurnal Sains dan Teknol.)*, vol. 8, no. 1, p. 12, 2019, doi: 10.23887/jst-undiksha.v8i1.12081.
- [15] B. Laugwitz, T. Held, and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire," *Lncs*, vol. 5298, p. 2007, 2007, [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.472.3719&rep=rep1&type=pdf>.
- [16] W. A. Pramono, H. M. Az-zahra, and R. I. Rokhmawati, "Evaluasi Usability pada Aplikasi MyTelkomsel dengan Menggunakan Metode Usability Testing," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2951–2959, 2019.
- [17] K. Ishaq, N. A. M. Zin, F. Rosdi, A. Abid, and Q. Ali, "Usability of mobile assisted language learning app," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 1, pp. 354–363, 2020, doi: 10.14569/ijacsa.2020.0110145.
- [18] G. Intan, "Bagaimana Kehidupan Pasca Pandemi Covid-19?," *Voice of America*, 2020. <https://www.voaindonesia.com/a/bagaimana-kehidupan-pasca-pandemi-covid-19/5393460.html> (accessed Mar. 01, 2021).
- [19] World Health Organization, "New Normal," *who.int*, 2020. <https://www.who.int/indonesia/news/novel-coronavirus/new-infographics/new-normal>.
- [20] kemsos.go.id, "What is New Normal?," *Kementrian Sosial Republik Indonesia*, 2020. <https://www.kemsos.go.id/en/what-is-new-normal>.
- [21] A. Schade, "Remote Usability Tests: Moderated and Unmoderated," Nielsen Norman Group, 2013. <https://www.nngroup.com/articles/remote-usability-tests/> (accessed Jan. 26, 2022).
- [22] N. Babich, "Usability Testing: Moderated vs Unmoderated," *medium.com*, 2020. <https://medium.com/thinking-design/usability-testing-moderated-vs-unmoderated-adbccc37404b> (accessed Jul. 09, 2021).
- [23] K. Whitenon, "Tools for Unmoderated Usability Testing," Nielsen Norman Group, 2019. <https://www.nngroup.com/articles/unmoderated-user-testing-tools/> (accessed Jan. 26, 2022).
- [24] J. M. C. Bastien, "Usability testing: a review of some methodological and technical aspects of the method," *Int. J. Med. Inform.*, vol. 79, no. 4, 2010, doi: 10.1016/j.ijmedinf.2008.12.004.
- [25] Minister of Education and Culture of the Republic of Indonesia, "Regulation of the Minister of Education and Culture of the Republic of Indonesia Number 2 of 2020 Concerning Procedures for Implementing Competency Test for Health Students," 2020.
- [26] RISTEKDIKTI, "Data Statistik UKNI," 2021. <http://ukners.ristekdikti.go.id/statistik> (accessed Jan. 21, 2019).
- [27] L. F. Manalu, "Menyikapi Krisis Kekurangan Perawat," *SINDOnews*, 2017. <https://nasional.sindonews.com/read/1210047/18/menyikapi-krisis-kekurangan-perawat-1496348683> (accessed Jan. 23, 2018).
- [28] *who.int*, "Rencana Pengembangan Tenaga Kesehatan Tahun 2011-2025," 2011.
- [29] K. Rajaguguk, "Indonesia Kekurangan Jumlah Tenaga Perawat," *Media Group*, 2020. <https://mediaindonesia.com/humaniora/314345/indonesia-kekurangan-jumlah-tenaga-perawat> (accessed Feb. 02, 2021).
- [30] S. Kholifah and W. Kusumawati, "Hambatan Lulusan Ners Dalam Menghadapi Uji Kompetensi Ners Indonesia," *Indones. J. Heal. Sci.*, vol. 7, no. 1, pp. 40–47, 2016.
- [31] A. Abdillah, "Analisis Faktor-Faktor Yang Mempengaruhi Kelulusan Uji Kompetensi Ners Indonesia," *J. Penelit. Adm. Publik*, vol. 2, no. 2, pp. 373–380, 2016, doi: 10.1038/nature07345.
- [32] L. Hakim and L. S. Pusporini, "Analisis Faktor Yang Mempengaruhi Capaian Kelulusan Uji Kompetensi Ners Mahasiswa Program Profesi Ners," *Cakrawala Pendidik.*, vol. 37, no. 2, 2018, doi: 10.21831/cp.v37i2.19881.
- [33] A. Hartina, T. Tahir, N. Nurdin, and M. Djafar, "Faktor Yang Berhubungan Dengan Kelulusan Uji Kompetensi Ners Indonesia (Ukni) Di Regional Sulawesi," *J. Persat. Perawat Nas. Indones.*, vol. 2, no. 2, p. 65, 2018, doi: 10.32419/jppni.v2i2.84.
- [34] H. A. Alzahrani and R. A. Alnanih, "A Design Study to Improve user Experience of a Procedure Booking Software in Healthcare," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 11, pp. 245–254, 2020, doi: 10.14569/IJACSA.2020.0111132.
- [35] K. Moran and K. Pernice, "Remote Moderated Usability Tests: Why to Do Them," Nielsen Norman Group, 2020. <https://www.nngroup.com/articles/moderated-remote-usability-test-why/> (accessed Jan. 26, 2022).
- [36] Sudjana, *Metode Statistika*. Bandung: Tarsito, 2002.
- [37] A. Al-Hunaiyyan, R. Alhajri, B. Alghannam, and A. Al-Shaher, "Student Information System: Investigating User Experience (UX)," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 2, pp. 80–87, 2021, doi: 10.14569/IJACSA.2021.0120210.
- [38] M. S. Bashir and A. Farooq, "EUHSA: Extending Usability Heuristics for Smartphone Application," *IEEE Access*, vol. 7, pp. 100838–100859, 2019, doi: 10.1109/ACCESS.2019.2923720.
- [39] J. Nielsen, "Usability Engineering," Nielsen Norman Group, 1993. <https://www.nngroup.com/books/usability-engineering/> (accessed Jan. 02, 2022).

4PCDT: A Quantifiable Parameter-based Framework for Academic Software Project Management

Vikas S. Chomal¹, Jatinderkumar R. Saini^{2*}, Hema Gaikwad³, Ketan Kotecha⁴

School of Engineering, P P Savani University, Kosamba, India¹

Symbiosis Institute of Computer Studies and Research, Symbiosis International (Deemed University), Pune, India^{2,3}

Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune, India⁴

Abstract—Many authorities like Project Management Body of Knowledge (PMBOK) and Capability Maturity Model Integration for Development (CMMI-Dev) lend a hand to software development organizations in management of their crucial projects. Though this area needs focused research, such models are not dedicatedly available for the academic projects developed by students of computer science and engineering where software project development is considered as one of the criteria for the award of degree to the future professionals of the IT industry. With this motivation, we explored 4PTRB, 3PR and software project management practices, approaches as well processes framed and provided by PMBOK and CMMI-Dev. The main aim of this research is to introduce and propose a software project management framework for the academic domain. The proposed framework contains identification and description of 7 and 26 quantifiable parameters and sub-parameters respectively. The framework is called 4PCDT for People, Process, Product, Project, Complexity, Duration and Technology for the academic software projects. To validate the proposed framework, an online survey of 113 faculties was conducted to rank and weigh the quantifiable parameters. The results show that People, Process and Technology management parameters are top 3 ranked parameters. The robustness of the approach is further evident from the results of experimentation on 18 actual academic software projects of final year post graduate students of the IT domain. Not only the proposed work is first of its kind, but also it is bound to generate an excellent ripple effect in the research community.

Keywords—Academic; CMMI-Dev; PMBOK; project management; software project; student

I. INTRODUCTION

Software project management plays a critical role in software project development. To manage project efficiently is considered as an art as well as a major demanding task in the Information Technology (IT) domain. The art and challenging role make both project development as well as management an extremely imperative research object in society. Project management has become a key process area as well as given a due importance irrespective of industry domain. To effectively deal with and manage project is considered as one of the decisive success causes for any software project [11]. Today software project management is deemed to have equivalent significance as those of applying software engineering concepts in software development environment and these both are considered as driving aspects to deliver a successful as well a qualitative software project outcome [15].

Software project development also has a significant importance in academic courses of computer science and engineering as well as information technology. Through this project development, practical knowledge of software development is imparted to students. During academic software project development students are provided with strict guidelines as well as instructed to mandatorily execute all the phases of system development life cycle [5][6]. One of the universal observations found in academic software project development is failure of students to accomplish software project development within a predetermined timeframe. There may be numerous causes behind this failure, but the most vital cause is inappropriate management of software project by students. It is extremely imperative to make students acquainted with guiding principles regarding proper software project management from the very preliminary period of software project development in order to endow them with intellectual IT proficiency.

The intention of study is: (a) to explore and evaluate software project management in software project development; (b) to indicate the significance of software project management in academic framework; (c) to propose an comprehensive methodology and framework that assists in managing software project considering academic context on the basis of rich literature review on software project management to encourage student's software project success probability; (d) to introduce a software project management framework by examining and incorporating earlier project management methodology/ frameworks that will guide, assist project mentors and students to deliver successful software project. The layout of the paper is as follows: a concise literature evaluation is discussed in Section 2. The planned framework is introduced in Section 3. Experiment and Result of the anticipated framework is presented in Section 4. Lastly conclusion and future enhancement drawn from this research is highlighted in Section 5.

II. LITERATURE REVIEW

According to Chang et al. [1] software project management is a problem-solving activity and task like other activities that are involved in software development process. Further, Chang et al. [1] also proposed a software project management model termed as SPMNet for resource allocation, scheduling and to track and handle project status. Marinho et al. [11] in their research work focused on various uncertainties that can be effectively managed using project management

*Corresponding Author.

techniques. The major contribution provided by Marinho et al. [11] are – (a) a systematic review for academic community to clearly understand about various challenges and uncertainties in project management, (b) techniques and strategies to deal with these uncertainties. Cristobal et al. [9] summarized project complexities and also discussed how to address these complexities using effective project management. Sajad et al. [15] define software project management as a process which starts from proper planning and then flow towards organizing, staffing, monitoring, controlling and leading a software project. He has presented a comparative analysis of various project management tools and also predicted about project management tools that will have greater impact on software development and quality. Mac and Pinto [10] stated that software project management has become a focused discipline in software engineering domain. Further, risk management was the major factor on which they proposed their findings in consideration with software project management. According to Mahdi et al. [14] planning and assessment are important activities of software project management and are considered to have immense effect on project performance and its outcome. Author's also presented an in-depth review on use of various machine learning algorithms in software project management. Cunha et al. [8] conducted a methodical literature evaluation on software project management and concluded through empirical study that decision making is one of the most important criteria in software project management and stated that more studies are needed to carry out to understand decision making fact from this naturalistic perception. Barghothet al. [13] affirmed that project management plays a central role for making software project a success story. They proposed a framework named 4PTRB which consist people, process, product technology, risk, and business management areas. The said methodology provides a complete and exhaustive support to software project administrators to get better their project administrating managing skills and efficacy.

In their research work, Alok and Deepti [12] focused that software processes applied as well as software project management are having due weight-age for developing a qualitative software. Also, they presented a comparative study of various project management tools that can be utilized for effective management of software development activities. Varajaoa [7] asserts that project management is discipline that has achieved a notable identification in research domain. Also, author states that irrespective of industries relevance of good project management practices leads to a successful project. Mira and Pinnington [3] in their research work tested the association between project management performance and project success and for their investigation they considered empirical data of project management professionals. Rehman and Hussain [20] reported exhaustive study on different project management methodologies and their importance in project management. Also, authors examined and presented a parameterize comparison between various project management methodologies with PMBOK.

Dey et al. [17] explored and described contemporary drift in software project management. Authors in their work analyzed and highlighted all categories of risks that are related

with technology, financial, scheduling, legitimate, principle, operational, security, communication, project and personnel as well as all these risks require timely involvement as well as proper follow-up and controlling needed in project management. The research work of Nakigudde [16] focused on foremost decisive factors that lead to the success as well as failure of the project. Author also explained the significant role played by project management model in making software project a successful journey. Demir et al. [2] examined and presented diverse approaches to examine the efficiency of project management in software development life cycle. Singh and Lano [19] worked out on techniques and framework of project management and their finding states that different techniques are suited and can be applied in different types of software project development. Kwak [21] scrutinized and presented in-depth history of project management as well enhancements that have taken place in domain of project management. Raj and Sinha [18] provided proposal on handling as well as enrichment in project management considering agile framework. Packer et al. [4] proposed a model that provides support in project management decisions in agile development considering the issues and difficulties faced in using GitHub repositories.

III. PROPOSED FRAMEWORK

For the present research work, a study and examination of previous project management models such as Project Management Body of Knowledge (PMBOK), Capability Maturity Model Integration - Development (CMMI-Dev) processes and 4PTRB (People, Process, Product, Project, Technology, Risk, and Business) [13] was carried out.

These project management models provide with a set of software project management approaches, procedures as well as directive philosophy for software project management discipline. Software project management framework 4PTRB [13] is considered as fundamental base model for research work. Since the said proposed framework is been implemented in academic context we borrowed People, Process, Product, Project and Technology software management areas from 4PTRB [13] and instead of software management areas we coined and define it as software management parameters. Parameters namely risk and business of 4PTRB [13] were not considered since our proposed framework is for educational project hence no risk factors need to be examined and evaluated similarly educational project are not developed considering business and other profit earnings. Also, these software management parameters are termed as quantifiable parameters since they are considered as a metrics for measuring software project management. Furthermore, two software management parameters are integrated to the existing 4PTRB [13] model framework.

One of the software management parameters concerned with academic domain is duration and the other one is complexity. The reason behind considering duration parameter is that academic software projects need to be completed within the stipulated time duration. Similarly, project mentors and students both have to examine and consider the complexity level of the software to be developed. In Table I, we present listing of identified software management parameters of the

proposed framework along with concise clarification of each parameter.

TABLE I. THE PROPOSED FRAMEWORK 4PCDT PARAMETER(S)

Sr. No.	Parameters	Depiction
1.	People	People are considered as one of the most important components of a project. Some of the assigned role in academic software project is team members and mentors.
2.	Process	Process is the clearly and well-defined roadmap that needs to be followed during software project development. In academic software development, students are strictly bound to follow defined process methodology.
3.	Product	Product refers to the outcome of the project, the main objective of the project. The students (team member) need to explain the product scope to the mentors and concerned authority so that the end results are understood to all the stakeholders.
4.	Project	The next parameter but not the least component is the project. This is where the huge role and accountability of the team members and mentors are under the limelight. The students need to execute as well as handle major development task as well as to ascertain timely completion of the phases and functionality of the software project development. Whereas, mentors have the task of overseeing the project, guiding and assisting team members with issues, and trying to ensure the project stays on track with the well-defined deadlines.
5.	Complexity	Complexity of software to be developed
6.	Duration	Stipulated time duration for completion of software project development
7.	Technology	Technology used for developing software

After preparing software project management areas list, next step is to identify and map sub-parameters for individual and main software project management parameters. 4PTRB software project management model [13] contains 28 sub areas. Further we revised the sub areas for the said proposed framework considering relevance and importance in academic software development and the same which is presented in Table II.

As revealed in Table II there are total seven main parameters and 26 sub-parameters for the proposed framework. The comparative analysis of main parameter(s) and sub-parameter(s) of proposed framework and 4PTRB [13] are summarized in Tables III and IV.

The primary objective of the research is to measure academic software project management efficiency based on the software project management parameters introduced in the anticipated framework. Therefore, a formula for measuring

project management effectiveness namely Academic Software Project Management efficacy (ASPME) is been introduced and the formula consists of the summing up of each main quantifiable parameters of software management. The formula for Academic Software Project Management Effectiveness (ASPME) is mentioned below [1]:

$$\text{ASPME Score} = \text{PeoplePW} + \text{ProcessPW} + \text{ProductPW} + \text{ProjectPW} + \text{ComplexityPW} + \text{DurationPW} + \text{TechnologyPW}$$

Here, ASPME Score = Academic Software Project Management Effectiveness Score, PeoplePW = People Parameter Weight, ProcessPW = Process Parameter Weight, ProductPW = Product Parameter Weight, ProjectPW = Project Parameter Weight, ComplexityPW = Complexity Parameter Weight, DurationPW = Duration Parameter Weight and TechnologyPW = Technology Parameter Weight.

Also these seven quantifiable parameters are not having equal weight-age. An online survey has been executed to endow with a rating to these academic software project management parameters. In the next section, the validation of the proposed framework including survey results and experimentation is presented.

TABLE II. IDENTIFICATION AND LISTING OF SUB-PARAMETERS

Sr. No.	Quantifiable Parameters	Proposed Framework (Sub-Parameters)	Sub Parameters Total
1.	People	Communication, Co-ordination, Team, Mentor and Team work	5
2.	Process	Project Identification, Project Feasibility, Project Planning, Project Monitoring & Controlling and Project Development Guidelines	5
3.	Product	Phase/Task verification & validation and Quality assurance	2
4.	Project	Phase/Task Definition, Phase/Task Allocation, Requirement Management, Reporting and Change Management	5
5.	Complexity	Project Domain, Project Scope, Team Size	3
6.	Duration	Task Duration Estimation, Monitoring & Controlling Task Duration and Verification & Validation of Task Completion	3
7.	Technology	Identification of Technology, Team Skills and Expertise and Knowledge Management	3
Total	7		26

TABLE III. COMPARISON OF MAIN PARAMETER(S) OF 4PCDT WITH 4PTRB [13]

Sr. No.	Software Project Management Model	Main Parameter	Total
1.	4PCDT	People, Process, Product, Project, Complexity, Duration and Technology	7
2.	4PTRB [11]	People, Process, Product, Project, Technology, Risk and Business	7

TABLE IV. COMPARISON OF SUB-PARAMETER(S) OF 4PCDT WITH 4PTRB [13]

Sr. No.	Main Parameter	Sub-Parameters		Total of Sub-Parameters	
		4PCDT	4PTRB [13]	4PCDT	4PTRB [13]
1.	People	Communication, Co-ordination, Team, Mentor and Team work	Communication, Teamwork, Leadership, Organizational Commitment, Project Manager, Stakeholder involvement, Staffing and Hiring	5	7
2.	Process	Project Identification, Project Feasibility, Project Planning, Project Monitoring & Controlling and Project Development Guidelines	Requirement Management, Project Planning, Project Monitoring & Control and Scope Management	5	4
3.	Product	Phase/Task verification & validation and Quality assurance	Configuration Management and Quality Engineering	2	2
4.	Project	Phase/Task Definition, Phase/Task Allocation, Requirement Management, Reporting and Change Management	Activity Definition, Activity Sequencing, Activity Resource Estimates, Activity Duration Estimates, Schedule Variance, Estimate Costs, Determine Budget and Cost Variance	5	8
5.	Technology	Identification of Technology, Team Skills & Expertise and Knowledge Management	Technology Maturation & Risk Reduction and Knowledge Management	3	2
6.	Complexity	Project Domain, Project Scope, Team Size	--	3	--
	Risk	--	Risk Management and Risk Control	--	2
7.	Duration	Task Duration Estimation, Monitoring & Controlling Task Duration and Verification & Validation of Task Completion	--	3	--
	Business	--	Contracting Management, Procurement Management and Benefit Management	--	3
Total				26	28

IV. EXPERIMENTAL RESULTS

For simplicity and enhanced inclusive research, the phased process was followed. These phases are presented below:

- Execution and analysis of online survey for assigning weights to parameters.
- Weight calculation for each parameter.
- To conduct experiment on data set.
- Analysis of the experiment result.

A. Execution and Analysis of Online Survey

Further, after identifying and listing these quantifiable parameters and sub-parameters attributes next procedure is to assign weights to these seven quantifiable parameters. For assignment of weights, we randomly selected one quantifiable parameter to begin with and going on to other parameters while keep on comparing the already assigned weights and the parameters to which weights are to be assigned.

This procedure was acknowledged by conducting an online survey for assigning weights to 7 quantifiable parameters by 113 faculties engaged in post graduate streams of information technology as well as computer science and engineering. These faculty members are having more than 10 years of academic experience as well as providing mentorship to students in their software project development. The averaged based on the values provided by 113 faculties are mentioned in tabular format in Table V.

It is significant to declare that each of the 7 quantifiable parameters were assigned weight out of 100 and it was not necessary to have the total of weights of 7 parameters as

break-up of 100. This course of action in principle is based on human perception and general aptitude.

B. Weight Calculation for each Parameter

In the next phase of research weights need to be calculated for each quantifiable parameter. The procedure implemented for the same is to divide average weight of each quantifiable parameter by total weight average as shown in Table VI.

TABLE V. AVERAGED VALUE OF QUANTIFIABLE PARAMETERS

Sr. No.	Quantifiable Parameters	Average (%)
1.	People	82.10
2.	Process	80.13
3.	Product	63.90
4.	Project	69.12
5.	Complexity	65.98
6.	Duration	70.04
7.	Technology	73.09

TABLE VI. WEIGHT AVERAGE TO QUANTIFIABLE PARAMETERS

Sr.No.	Quantifiable Parameters	Average (%)	Weight
1.	People	82.10	0.1621
2.	Process	80.13	0.1582
3.	Product	63.90	0.1261
4.	Project	69.12	0.1365
5.	Complexity	65.98	0.1308
6.	Duration	70.04	0.1383
7.	Technology	73.09	0.1443
Total	7	506.36	1.0000

TABLE VII. SIGNIFICANCE OF 4PCDT

Results	Choices						Total
	Very Significant	Significant	Somewhat Significant	Neutral	Not Significant	No Opinion	
Number of Respondent(s)	61	31	13	5	3	0	113
Percentage (%)	53.98	27.43	11.50	4.42	2.65	0	100

In the online survey form given to the respondents, the respondents were also asked about the significance of the proposed framework called 4PCDT. The respondents were informed that this framework could be employed as a guiding principle for the software project management for the academic domain. The results were analyzed, summarized and presented in Table VII.

C. Experiment on Dataset

The perform experiment on dataset is an imperative element of the research. Similarly, the practical execution of the proposed framework was executed in our organization. For experiment, software project developed by Master Degree students were considered. Final year students need to develop this academic software project within six months. We inspected and assessed 18 large academic software projects developed during the three consecutive years 2016-2017, 2017 – 2018 and 2018-2019 and led by 5 faculties and there mentored the same software projects. Further free online Software Project Management Effectiveness Evaluator (SPMEE) tool named Wrike was used to perform the experiment of the proposed framework. This tool provides with a facility where by we can design self-administered project management effectiveness questionnaire. A structured and organized set of closed-form questionnaire was prepared considering academic software project development life cycle based on the implementation of the proposed framework taking into consideration quantifiable parameters introduced.

D. Analysis of Experimental Result

The academic software projects considered for experiments are presented in Table VIII. Further, all faculty participants were provided with project management effectiveness questionnaire. In the next step, academic project success scores need to be provided by each faculty members using this online software project management efficacy evaluator and project management effectiveness scores calculated were in-between range 0 to 10. Where 0 means a software project is not successful stating that least effective project management parameters have been functionally applied by students and mentors. While a score 10 denotes an extremely successful academic software project were at most care is taken as well as foremost efficient software project

administration and execution has been functionally implemented.

Finally, we compared the results obtained by us through the 4PCT model, which is the modified 4PTRB model with 4PTRB [13] project management model itself. The implementation was done on the said 18 academic software projects and the software project management effectiveness was measured. Each academic software projects were solely varied from other projects in the dataset. The development time duration for each academic software projects was 6 months whereas each project varies in domain, functionality, team size, technology, complexity. In Table IX in-depth experiment result analysis and comparison of proposed framework and 4PTRB [13] is presented. Software project management score is automatically measured by the online Software Project Management Effectiveness Evaluator (SPMEE) considering the Academic Software Project Management Effectiveness (ASPME). These analysis and findings strengthen the legitimacy of the proposed framework.

It has been found that project success score and software project management effectiveness (PME) are closely associated with each other and have a strong correlation. The same is graphically presented in Fig. 1. Also the association between proposed framework PME score and success score is stronger than 4PTRB [13] PME score and success score. Further, the Pearson correlation co-efficient is 0.9754, while it is 0.9288 when the 4PTRB [13] framework is applied. Thus, it can be wrapped up that employing the proposed framework highlights the higher probability of delivering as well as managing software project more effectively and successfully. Also, it can be observed that score generated by proposed framework 4PCDT and Success rate score is closer in comparison to 4PTRB [13].

TABLE VIII. ACADEMIC SOFTWARE PROJECT DEVELOPMENT YEAR

Sr. No.	Academic Year	Project Considered for Experiment
1.	2016-2017	6
2.	2017-2018	6
3.	2018-2019	6
	Total	18

TABLE IX. ANALYSIS OF DATASET

Sr. No.	Project Title	Development Year	Team Size	Project Completion(%)	Project Type	Success Rate	4PCDT (Proposed)	4PTRB [13]
1	APMC Mgt System	2016-2017	2	91	Desktop	6	6.12	5.11
2	E-Shop	2016-2017	2	100	Web-Based	8	8.09	7.52
3	Online Multistore Portal	2016-2017	2	80	Web-Based	6	6.23	6.12
4	E – Library	2016-2017	1	90	Web-Based	7	7.02	6.08
5	On line Exam	2016-2017	2	100	Mobile App	8	8.01	7.71
6	Online Shopping Portal	2016-2017	3	100	Mobile App	7	7.67	6.62
7	Corporate Recruitment Mgt System	2017-2018	3	90	Web-Based	7	7.31	6.19
8	Online Review System	2017-2018	2	100	Mobile App	8	8.05	7.81
9	Restaurant Mgt System	2017-2018	2	100	Web-Based	7	7.81	6.02
10	College Mgt System	2017-2018	2	95	Mobile App	6	6.46	5.07
11	Work Flow Mgt System	2017-2018	1	90	Desktop	6	6.21	5.82
12	Rental Application	2017-2018	1	80	Web-Based	5	5.09	5.03
13	Car Pooling System	2018-2019	2	100	Mobile App	7	7.11	6.21
14	Inventory & Supply Chain Mgt System	2018-2019	3	100	Web-Based	7	7.19	6.52
15	Digital Campus	2018-2019	2	100	Web-Based	8	8.41	7.12
16	Milk Distribution	2018-2019	1	80	Desktop	6	6.36	5.09
17	Billing System	2018-2019	2	100	Desktop	8	8.11	6.86
18	Production Monitoring System	2018-2019	3	80	Web-Based	5	5.09	4.19

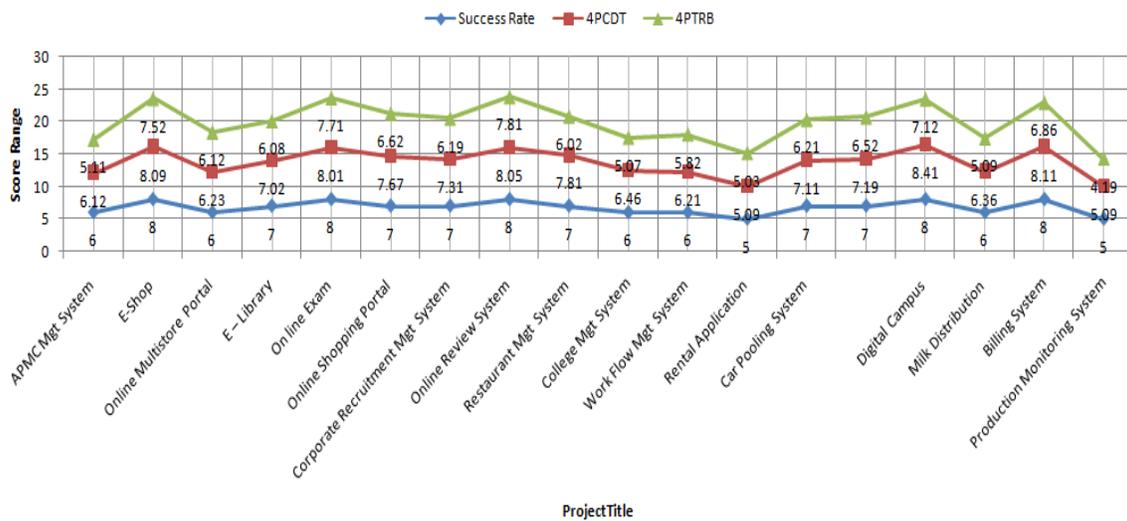


Fig. 1. Relationship between the Project Success Rate and 4PCDT as well as 4PTRB [13] PME score.

V. CONCLUSION AND FUTURE ENHANCEMENT

In the present research we proposed an academic software project management framework named as 4PCDT which is developed with consideration of multiple parameters. Seven main quantifiable parameters and 26 sub-parameters were recognized and listed based on their relevance. The primary aim of categorizing these parameters was to provide with the academic software project management framework which is best suited in consideration with academic context. Existing software project management frameworks like 4PTRB [13],

3PR and various guidelines were explored for proposing the academic software project management framework. However, the proposed framework is having unique characteristics because we revised the parameters and sub-parameters which were included in earlier versions of project software management framework and introduced list of new parameters and sub-parameters to cover more facets and propose a more holistic and comprehensive framework for managing academic software projects. Validation of the identified quantifiable parameters and sub-parameters was done through more than 100 faculties of post graduate courses of computer

science, computer engineering and information technology. All chosen faculties were active mentors for academic software project development. The results showed that 'People' management has been considered with maximum significance followed by 'Process' and 'Technology' in the academic software project management domain.

In the present work 18 academic software projects were used to experiment and validate the proposed framework. The academic project development work was carried out during the academic years 2016-2017 to 2018-2019. In the next step of research we prepared project management questionnaire and was provided to five faculty participants. Faculty members used this questionnaire on these 18 academic software projects and provided project success scores, project management effectiveness score using online Software Project Management Effectiveness Evaluator (SPMEE) tool. Further, scores were calculated considering the range 0 to 10, where 0 signifies that a software project is not successful and the cause behind this is least effective project management parameters has been practiced by students and mentors. While a score 10 means an extremely successful academic software project where at most care is taken as well as foremost effective software project management has been functionally implemented. Same technique was considered in the previous framework and their studies stated a positive association and relationship between software project success score and project management effectiveness.

Similarly, the finding and analysis of present research shows a strong and optimistic interrelationship between software project success score and project management efficiency with 0.9754 value of Pearson correlation coefficient whereas it is 0.9288 when 4PTRB [13] framework is applied. Thus, it can be concluded by findings of the analysis that the proposed framework is hypothetical, optimum, applicable and appropriate to be used in academic software project management. Considering the same we deem that academic courses that are having major as well as minor software project development as a part of their core curriculum should emphasize, consider and endow with course of action as well as models and methodologies regarding software project management in software project development.

The extensive framework presented through this research work will definitely assist the faculty mentors as well as the students in the domains like Information Technology, Computer Science, Computer Engineering, and Computer Application to manage the academic software development projects more effectively. Proceeding with the research, we would execute the work in direction to introduce software project management effectiveness model for academic domain in consideration with proposed framework. In the current research work only 18 academic software projects were included in the experiment; hence if the size of dataset for validating the framework is increased it may disclose novel dimensions. Also, the proposed framework considered the academic context whereas 4PTRB [13] framework was designed considering software projects developed in IT industries. Hence at last, we express that the proposed framework for academic software project management is

unconditional independent, reliable, prescribed as well as shows a better participation of students and faculty mentors and can be effortlessly employed in academic outline irrespective of project categories.

REFERENCES

- [1] Carl K. Chang, Chikuang Chao, Thinh T. Nguyen Mark Christensen, "Software Project Management Net: A New Methodology on Software Management", Proceedings. The Twenty-Second Annual International Computer Software and Applications Conference, 2002, IEEE. doi: 10.1109/CMPSAC.1998.716715.
- [2] Demir, K.A., Michael, J.B. and Osmundson, J.S, "Approaches for Measuring Software Project Management Effectiveness", International Conference on Software Engineering Research and Practice, Las Vegas, Vol. 2, 613-619, 2009. doi: 10.21236/ADA484712.
- [3] Farzana Asad Mira, Ashly H. Pinningtonb, "Exploring the Value of Project Management: Linking Project Management Performance and Project Success", International Journal of Project Management, 32, 202-217, 2017. doi: 10.1016/j.ijproman.2013.05.012.
- [4] Heather S. Packer, Adriane P. Chapman, L. Carr, "GitHub2PROV: Provenance for Supporting Software Project Management", Published in TaPP 2019 Computer Science, Engineering, 2019.
- [5] J.R. Saini, V.S. Chomal, "SaiCho: A Parameters Based Model for Team Building for Academic Software Projects", proc. of IEEE Inter. Conf. on Electrical, Computer and Communication Technologies (ICECCT-2015), Coimbatore, India; Feb. 2017, pp. 1129-1138 3.
- [6] J.R. Saini, V.S. Chomal, "Domain-based Ranking of Software Test-effort Estimation Techniques for Academic Projects", proc. of Inter. Conf. on ICT for Sustainable Development (ICT4SD-2019), Panaji, India; in press with AISC, Springer, Mar. 2020.
- [7] Joao Varajao, "Success Management as a PM knowledge area – work-in-progress", Procedia Computer Science 100 (2016) 1095 – 1102, Available online at www.sciencedirect.com, 2016.
- [8] Jose Adson O. G Cunha, Hermano P. Moura, Francisco J.S.Vasconcellos, "Decision-Making in Software Project Management: A Systematic Literature Review", Procedia Computer Science 100 (2016) 947-954, 2016.
- [9] Jose R. San Cristobal ,Luis Carral , Emma Diaz, Jose A. Fragueta and Gregorio Iglesias, "Complexity and Project Management: A General Overview", Review Article, Hindawi Complexity, Volume 2018, Article ID 4891286, 10 pages, 2018. doi: 10.1155/2018/4891286.
- [10] Kevin MacG, C. Ariel Pinto, "Software Development Project Risk Management: A Literature Review", 26th ASEM National Conference Proceedings, October 2005.
- [11] Marcelo Marinho, Suzana Sampaio, Telma Lima and Hermano de Moura, "A Systematic Review of Uncertainties in Software Project Management", International Journal of Software Engineering & Applications (IJSEA), Vol.5, No.6, 2014. doi: 10.5121/ijsea.2014.5601.
- [12] Mishra Alok, Mishra Deepti, " Software project management tools: a brief comparative view" , ACM SIGSOFT Software Engineering Notes, May 2013 Volume 38 Number 3, 2013. doi: 10.1145/2464526.2464537.
- [13] Mohamed Ellithy Barghot, Akram Salah, Manal A. Ismail, "A Comprehensive Software Project Management Framework", Journal of Computer and Communications, 2020, 8, 86-102 <https://www.scirp.org/journal/jcc>, ISSN Online: 2327-5227, ISSN Print: 2327-5219, DOI: 10.4236/2020.
- [14] Mohammed Najah Mahdi , Mohd Hazli Mohamed Zabil , Abdul Rahim Ahmad , Roslan Ismail , Yunus Yusoff , Lim Kok Cheng , Muhammad Sufyan Bin Mohd Azmi , Hayder Natiqand Hushalini Happala Naidu , "Software Project Management Using Machine Learning Technique—A Review", 8th International Conference on Information Technology and Multimedia (ICIMU), Publisher: IEEE, DOI: 10.1109/ICIMU49871.2020.9243543, 2020.
- [15] Muhammad Sajad, Muhammad Sadiq , "Software Project Management: Tools assessment, Comparison and suggestions for future development", IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.1, January 2016.

- [16] Nakigudde, S, "Project Management Models and Software Development Project Success", <https://doi.org/10.13140/RG.2.2.36203.08482>, 2019.
- [17] Pradip Peter Dey, Mohammad Amin, Bhaskar Raj Sinha, Shatha Jawad, Laith Al Any, Hassan Badkoobehi, "Current Trends in Software Project Management", *Advances in Social Science, Education and Humanities Research*, volume 120, DOI:10.2991/mshsd-17.2018.5, World Conference on Management Science and Human Social Development, 2017.
- [18] Pritraj, Parul Sinha, "Project Management In Era Of Agile And Devops Methodologies", *International Journal Of Scientific & Technology Research* Volume 9, Issue 01, January 2020.
- [19] Ravinder Singh, Kevin Lano, "Literature Survey of Previous Research Work in Models and Methodologies in Project Management", *International Journal of Advanced Computer Science and Applications*, Vol. 5, No. 9, 2014.
- [20] Ur Rehman, A. and Hussain, R, "Software Project Management Methodologies/ Frameworks Dynamics "A Comparative Approach". *International Conference on Information and Emerging Technologies* , DOI: 10.1109/ICIET.2007.4381330, Publisher: IEEE, 2007.
- [21] Young Hoon Kwak, "Brief History of Project Management", Chapter 2 in *The Story of Managing Projects* by Carayannis, Kwak, and Anbari (editors), Quorum Books, 2003.

Neuromarketing Solutions based on EEG Signal Analysis using Machine Learning

Asad Ullah, Gulsher Baloch, Ahmed Ali, Abdul Baseer Buriro, Junaid Ahmed, Bilal Ahmed, Saba Akhtar

Department of Electrical Engineering
Sukkur IBA University, Airport Road Sukkur
Sindh, Pakistan

Abstract—Marketing campaigns that promote and market various consumer products are a well-known strategy for increasing sales and market awareness. This simply means the profit of a manufacturing unit would increase. "Neuromarketing" refers to the use of unconscious mechanisms to determine customer preferences for decision-making and behavior prediction. In this work, a predictive modeling method is proposed for recognizing product consumer preferences to online (E-commerce) products as "Likes" and "Dislikes". Volunteers of various ages were exposed to a variety of consumer products, and their EEG signals and product preferences were recorded. Artificial Neural Networks and other classifiers such as Logistic Regression, Decision Tree Classifier, K-Nearest Neighbors, and Support Vector Machine were used to perform product-wise and subject-wise classification using a user-independent testing method. Though, the subject-wise classification results were relatively low with artificial neural networks (ANN) achieving 50.40 percent and k-Nearest Neighbors achieving 60.89 percent. Furthermore, the results of product-wise classification were relatively higher with 81.23 percent using Artificial Neural Networks and 80.38 percent using Support Vector Machine.

Keywords—*Electroencephalogram (EEG); brain-computer interface; neuromarketing; machine learning; artificial neural networks*

I. INTRODUCTION

E-commerce is a growing field these days. People want to expand their businesses, so they spend money on marketing to learn about their customers' preferences. Neuromarketing is a developing field with enormous potential for application marketing, brand management, and advertising. It emerges as a result of combining relevant concepts from the fields of neural science, psychology, human neurophysiology and even neuro chemistry. It connects consumer behaviour research with neuroscience [1]. Consumer behaviour quite often undermines the effectiveness of traditional marketing methods.

This is because the consumers' reactions vary when they are exposed to advertisements. Neuromarketing is the key to gaining insight into the minds of consumers. Because neuromarketing does not necessitate the consumer's conscious participation. It operates on the unconscious state of the brain.

Neuromarketing assesses the brain's reaction to any advertising stimuli. It differs from self-reports that consumers

provide during surveys. The truth can be revealed by studying the EEG signals directly [21]. As several reported studies show those two systems- conscious and subconscious can provide contradictory interpretations at times. Individual choices influence the decision-making process not only through individual and cognitive assessments, such as questionnaire responses but also through objective and implicit assessments, such as eye movement and neural activities. Recent findings from functional magnetic resonance imaging (fMRI) and EEG studies have linked movements in the frontal theta and posterior gamma bands to the development of individual choice. These findings show that before deliberate decision making, the physical reaction is caused by implicit desires. As a result, such neural behaviors associated with attention-related tasks, such as eye moments, can influence the consumer's preferences at an unconscious level. Despite this, there have been few neurological studies on the relationship between visual attention and subjective interest: the causes of subjective preference choices, such as the amount of visual perception and attention, are impossible to assess when using attractive faces with a wide range of visual features (e.g., facial contour, eye color, and hair length) [2].

Several commercial efficacy metrics can be calculated using neuromarketing. Emotional commitment, memory retention, purchase purpose, novelty, perception, and attention are the factors to consider. When customers make decisions, they are influenced by their emotions. The emotional interest level causes the emotional commitment level to rise. It can also help predict when customers will purchase by observing how their brains react to advertising stimuli. When customers decide to buy a product, the level of encoding of marketing stimuli influences our decision [3].

Neuromarketing provides knowledge that traditional marketing methods cannot provide. The significant advantage provided by neuromarketing techniques is that these techniques, which collect quantitative data, could be used before the launch of a new product, increasing the likelihood of that product's success [3].

Electroencephalography (EEG) was developed to record brain signals. EEG is used to study brain activity by recording postsynaptic potentials generated by neurons. With the development of tools, EEG is no longer limited to medical applications but has now been extended to other fields. Medical, Brain-Computer Interface (BCI), and neuromarketing are examples of EEG applications [4]. In the

[10-11], the authors have proposed a predictive modeling method based on EEG signals to understand customer preferences for E-commerce products in terms of "likes" and "dislikes". EEG signals were recorded while volunteers of various ages and gender browsed through various consumer goods. The tests were performed on a dataset containing a variety of consumer goods. The accuracy of choice prediction was calculated using a user-independent testing approach and hidden Markov Model (HMM) classifier. The prediction results appear to be promising, and the methodology can be used to create business models [11].

In comparison to the previous study, this study introduces subject-wise classification as well. The previous study has only done on the product-wise classification. The goal of this study is to assist marketing researchers in making appropriate decisions for further increasing the sale of products using imaging techniques by developing an EEG-enabled model that can replace expensive methods of current day neuromarketing. In addition, by analyzing EEG signals, a Neuro-marketing system will be provided to predict customer choices while viewing E-commerce products.

As such, the main objective of this study is to investigate the various tuning of artificial neural networks and other classifiers for improving the classification rates of product-wise classification and for the first time doing subject-wise classification. Section II presents the background and related works in the field of neuromarketing. Section III presents our approach towards building an EEG-based prediction model. Section IV presents the results of our study and Section V concludes the paper with possible future recommendations.

II. RELATED WORK

We looked at recent studies that linked EEG signals to predict customers' response, behavior and emotions based on self-reported ratings. All these studies mostly focus on studying the relationship between brain imaging and customer decision-making. Kumar, Singh, et al. (2015) investigated the current state of neuromarketing, as well as the activities involved, which included neuroimaging, EEG, fMRI, and eye-tracking. The customer dialectic is examined in the paper: "consumers contradict themselves, saying what they want but doing what they feel." The authors focused on four aspects of consumers: physical body, mind, heart, and spirit [5].

W. Anderson, Sijercic et al. (2007) worked on the classification of EEG Signals from four subjects while performing five mental tasks. Half-second segments of six-channel EEG data were trained to be graded into one of five groups, each of which corresponds to one of five cognitive tasks completed by four subjects. Two and three-layer feed forward neural networks were trained using 10-fold cross-validation and early stopping to avoid over fitting. To represent EEG signals, autoregressive (AR) models were used. The average percentage of correctly classified test segments ranged from 71% for one subject to 38% for another. The Clustering of the hidden-unit weight vectors of the resulting neural networks shows which EEG channels were most important in this discrimination problem [6, 20].

Solhjoo, Nasrabadi, Golpayegani, et al. (2005) investigated chaotic signal classification using hmm classifiers and EEG-based mental task classification. The analysis of mental activities using brain signals, based on EEG provides a better understanding of human brain functions. Furthermore, the author stated for EEG chaotic signals it is critical to determine whether probabilistic and statistical signal processing tools (such as HMM-based classifiers) can handle chaotic signals. The author has examined how well HMMs perform in classifying various types of synthetically formed chaotic signals. The performance of such classifiers in classifying mental tasks based on EEG was then evaluated. The results in both cases indicate good performance [7].

Guo et al. (2013) developed the new recommender system for 3D e-commerce using EEG signals. The author proposed a novel augmented reality recommender framework for the world of e-commerce. The system makes recommendations based on customer preferences, taking into account both pre-purchase and post-purchase scores, as well as post-purchase ratings in general. Positive emotions among users are evaluated using EEG signals before interacting with 3D virtual products. Pre purchase ratings work in tandem with post-purchase ratings to address two major challenges that traditional recommender systems face: data and cold start. By properly utilizing both pre-and post-purchase scores, user preference can be more reliably modeled. The author claimed that it has boosted the effectiveness of modern recommender systems and force traditional ecommerce applications to adapt [8].

The authors of [9] conducted an experiment on EEG signal classification using the wavelet transform. The author used an artificial neural network (ANN) technique in conjunction with a feature extraction technique, namely the wavelet transform. The artificial neural network used to classify the data is a feed-forward network with three layers that implements the back propagation algorithm for error learning. After that, the network with wavelet coefficients was trained. Over 66% of the normal class was correctly graded, and 71 % of EEGs in the schizophrenia group were positive.

Murugappan, Celestin Gerard et al. (2014), the goal of their research is to use wireless EEG signals to identify the most popular automotive brand in Malaysia. This work is taken into account a community of four major vehicle brand advertisements, including Toyota, Audi, Proton, and Suzuki. The participants (9 male and 3 female, ages 22-24) were simulated using a 14 channel wireless Emotive headset with a sampling frequency of 128 Hz, and the brain activity responses to the stimuli were obtained using a 14 channel wireless Emotive headset with a sampling frequency of 128 Hz. The obtained signals are filtered using a Surface Laplacian filter and a 4th order Butterworth band pass filter with a cut-off frequency of 0.5 Hz - 60 Hz is used to filter the obtained signals. The alpha frequency band (8 Hz - 13 Hz) of EEG signal information was obtained using the same Butterworth 4th order filter. The Fast Fourier Transform (FFT) was used to extract three statistical features from EEG signals using the Alpha band frequency spectrum: power spectral density (PSD), spectral energy (SE), and spectral centroid (SC). The feature vector is constructed using extracted features extracted

from all of the subjects via four different advertising stimuli. This feature vector is fed into two non-linear classifiers, K Nearest Neighbor (KNN) and Probabilistic Neural Network (PNN), to classify the subject's intention on advertising [10].

III. SYSTEM SETUP

In this study EEG signals were recorded from 15 healthy people using a Muse 2 headset – which is a neuro-signal acquisition wireless device – connected to a mobile app called Muse Monitor, as shown in Fig. 1. The device has 4 channels for EEG data that are located at AF7, AF8, TP9, and TP10 positions as per the International 10 - 20 system. Internally Muse 2 headband is sampled at a frequency of 256 Hz. The EEG data is stored in a CSV file and then transferred to computer for further processing. EEG headset was mounted onto the head of participants and asked to view shopping products as shown in Fig. 4.

We recorded 450 EEG signals, each lasting 4 seconds. Because the Muse 2 device has four sensors, it is the most user-friendly data acquisition device. We get raw data from four sensors. Fig. 2 shows the raw signal from AF7 Channel. Fig. 3 is the graph of RAW signals from AF7, AF8, TP9 and TP10 sensors. Fig. 4 shows the products we have used. While the user was viewing an item EEG signals were collected simultaneously. After the watching, each consumer was asked to rate the product in one of two categories: like or dislike. Then the signal passes through certain signal pre-processing techniques and feature extraction steps. Next, classification models are built, trained, and tested based on the user's choice.

A. Data Preprocessing and Feature Extraction

Pre-processing is the necessary step in EEG processing because it converts the signal into a usable format. The Initial pre-processing was done in excel to ensure that each recording was exactly 4 seconds long. Fig. 3 shows the unfiltered raw EEG signals from different channels: AF7, AF8, TP9 and TP10.

B. S-GOLAY Filter

Researchers have successfully used the S-Golay filter in signal smoothing. It is implemented using least squares or polynomials to reduce the noise in signals and smooth them out by fitting consecutive sub-sets of neighboring signal points with low-degree polynomials and linear least squares. The S-G filter has two parameters: polynomial degree and frame size. PRO and SNR (signal to noise ratio) is the output evaluating variables for denoising EEG signals using the S-G filter. The experiment results show which type of polynomial degree value is best [13].

The S-Golay filter can be applied to obtain a smoothed signal for a signal $S_j = f(t_j)$, where $(j=1, 2, \dots, n)$ with length of n as mathematically defined below.

$$Q_j = \sum_{i=-\frac{m-1}{2}}^{\frac{m-1}{2}} c_i S_{j+1}, \frac{m+1}{2} \leq j \leq n - \frac{m-1}{2} \quad (1)$$

Where m is defined as frame span, c_i represents convolutional coefficients number and Q_j is the smoothed output signal. Fig. 4 depicts the RAW signal from a single

channel, while Fig. 5 depicts the smooth signal after applying the S-Golay filter.



Fig. 1. Muse 2 Headband [12].

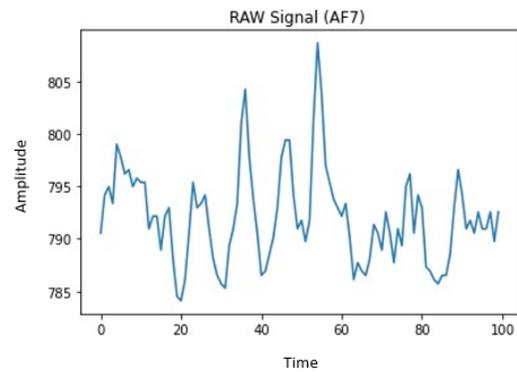


Fig. 2. Raw Signal from AF7 Channel.

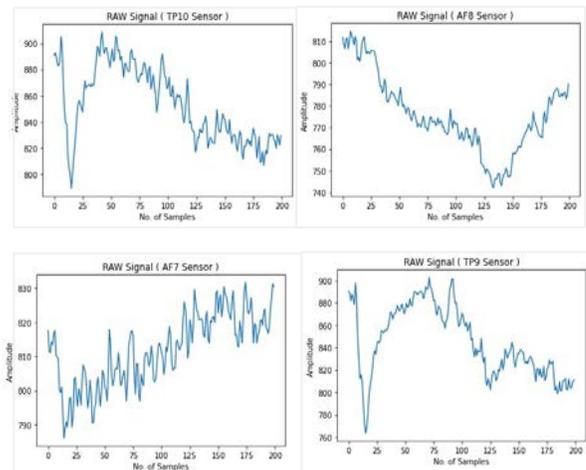


Fig. 3. Raw Signals from the EEG recording.

Item type	Sample 1	Sample 2	Sample 3	Item type	Sample 1	Sample 2	Sample 3
Shirts				Jeans			
Shoes				Formal Shirts			
Glasses				Joggers			
Watch				Bags			
Belts				Jackets			

Fig. 4. Product Images.

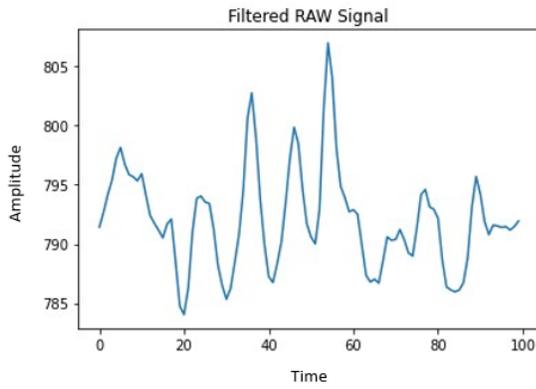


Fig. 5. Filtered Raw Signal.

C. Wavelet Transform Wavelet Transform (DWT) based Features

The most important part of distinguishing objects from one class to another is feature extraction. It is the process of converting raw signals into useful features. It is required to proceed to the next steps. For the classification of the EEG signal, we used discrete wavelet transform (DWT) [14] based on features. The DWT typically produces five signals: alpha, beta, theta, gamma, and delta bands of varying frequencies. Eq. 2 provides a mathematical explanation of DWT.

$$\phi(x) = \sum_{k=-\infty}^{\infty} a_k \phi(Sx - k) \quad (2)$$

When S is used as a scaling factor, it is usually set to 2. DWT is commonly used in biomedical signal processing because it denotes a signal in the time and frequency domains. The basic idea, behind DWT, is to use multistage decomposition to transform the signal input signals into small waves. A signal's wavelet analysis based on transformation can be performed at different frequency bands by decomposing it into approximation (A) and information (D) coefficients. To begin, two digital filters, the Low Pass Filter (L) and the High Pass Filter (H) are used to process the signal (H). A low-pass filter (L) is applied to the signal, which eliminates high-frequency fluctuations while preserving slow patterns.

D. Classification

Following the feature extraction step, we used those features for classification. Support Vector Machine (SVM) [15], Logistic Regression [16], Decision Tree [17], Random Forest [18], and Artificial Neural Networks [19] are among the classifiers employed. These features were classified subject-wise as well as product-wise on different bands. The dataset were divided into training and testing sets, with 80 percent of our data used for model training and 20 percent used for model testing.

IV. RESULTS AND DISCUSSION

As the features are fed into the Artificial Neural Network, various classifiers such as SVM, LDA, Logistic Regression, Random Forest, and Decision Tree are employed. The following are the ANN results.

A. Subject Wise Classification

We used 14 subjects in our experiment to collect EEG data from them while they are watching and selecting the products, and a K-fold cross validation of 10 folds was used to validate our experimental results. Different models are trained and accuracies are obtained using 5 different bands: alpha, beta, theta, gamma, and delta. These accuracies are evaluated for each product separately to carry out subject-wise classification.

B. Hyper Parameter Tuning on Theta Band for Model Selection

ANN is trained on 14 subjects using theta band with columns named Theta AF7, Theta AF8, Theta TP9, and Theta TP10.

Table I displays the tuning of the ANN model's model parameters. To achieve the best results, the hidden layers, the number of neurons, activation function on layers, and optimizer are all varied. As a result, the number of hidden layers should be one and the number of neurons should be two to achieve the best result of 50.40 percent for theta band.

ANN Model is trained using theta, alpha, beta, gamma, and delta bands using 10 folds on 14 subjects. Different parameters of ANN models have been tuned to achieve the best overall accuracy for each band.

Subject-wise accuracy of 50.40 percent, 50.02 percent, 50.39 percent, 50.14 percent, and 50.21 percent is obtained using Artificial Neural Networks on the Theta band, Alpha band, Beta band, Gamma band, and Delta band, as shown in Table II by testing different classifiers to obtain the best subject-wise accuracy. Using Decision Tree, K-nearest Neighbors, and Logistic Regression, we achieve the highest accuracy on the Delta band of 57.30 percent, 60.89 percent, and 51.34 percent, respectively. Hence, K-nearest neighbors proved to be best algorithm for classification of the Delta band signals.

Fig. 6 shows the accuracy of theta band tested number of times. Fig. 7 shows the maximum attained accuracy by alpha band is 51.5 %. Fig. 8 illustrates the minimum accuracy for beta band is 49.2 % and maximum accuracy is 50.5 %. Fig. 9 and Fig. 10 depict that minimum accuracy for delta and gamma band almost same that is 45.5 % with having maximum accuracy of 52 %.

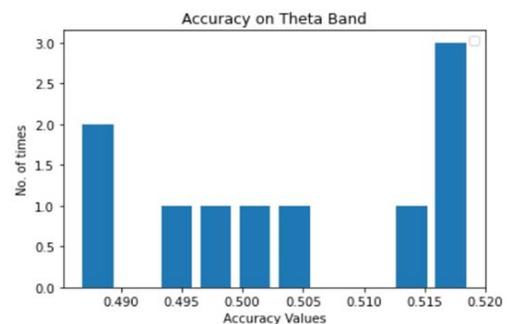


Fig. 6. Accuracies obtained on Theta Band.

TABLE I. SUBJECT-WISE ACCURACY TUNING FOR THETA BAND USING ANN

No of hidden layers	No of Neurons	Activation function on layers	Optimizer	Accuracy mean
1	2	[Relu, relu,Sigmoid]	Adam	49.85%
1	4	[Relu, relu,Sigmoid]	Adam	49.85%
1	8	[Relu, relu,Sigmoid]	Adam	49.93%
1	16	[Relu, relu,Sigmoid]	Adam	50.42%
1	32	[Relu, relu,Sigmoid]	Adam	50.24%
1	64	[Relu, relu,Sigmoid]	Adam	49.86%
1	128	[Relu, relu,Sigmoid]	Adam	50.21%
1	256	[Relu, relu,Sigmoid]	Adam	50.22%
1	512	[Relu, relu,Sigmoid]	Adam	49.11%
2	[2,4]	[Relu, relu,Sigmoid]	Adam	49.25%

TABLE II. SUBJECT-WISE ACCURACIES ON DIFFERENT BANDS USING ANN

Band	No. of hidden layers	No. of Neurons	Activation function on layers	Optimizer	Accuracy mean
Theta	1	2	[Relu, relu, sigmoid]	Adam	50.40%
Alpha	1	2	[Relu, relu, sigmoid]	Adam	50.02%
Beta	1	2	[Relu, relu, sigmoid]	Adam	50.02%
Delta	1	2	[Relu, relu, sigmoid]	Adam	50.21%
Gamma	1	2	[Relu, relu, sigmoid]	Adam	50.14%

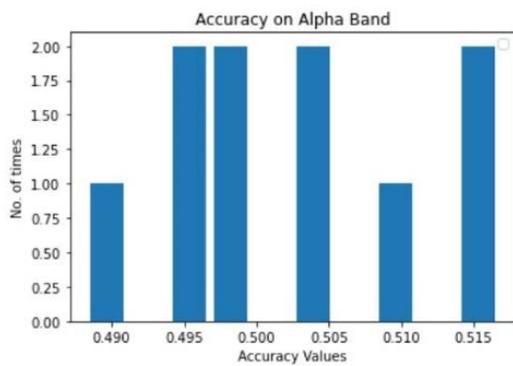


Fig. 7. Accuracies obtained on Alpha-band.

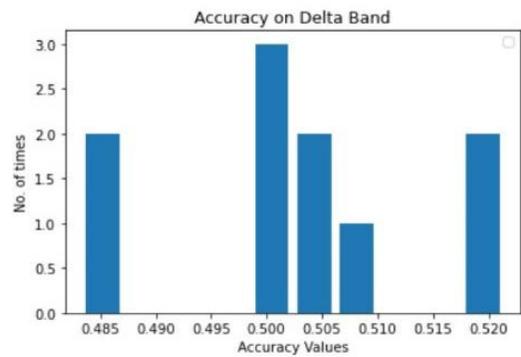


Fig. 9. Accuracies obtained on Delta band.

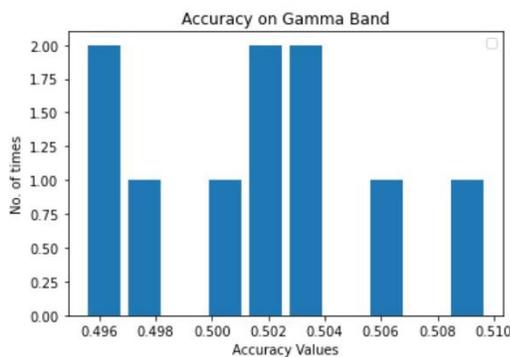


Fig. 8. Accuracies obtained on the Beta Band.

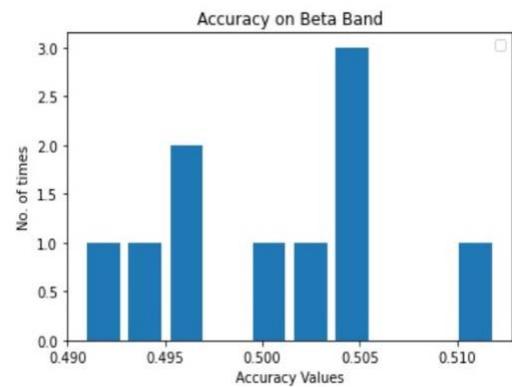


Fig. 10. Accuracies obtained on Gamma-band.

C. Product Wise Classification

The K-fold cross validation with 10 folds is applied for the product as well as subject-wise classification. Training and testing accuracies are obtained for various models using five different bands: alpha, beta, theta, gamma, and delta. Fig. 11 depicts the accuracy graph of the ANN model, which was trained using 14 products 1 subject. The model has learned completely after 15 epochs.

As shown in Table III, the average product-wise results are obtained accuracy of 78.73 percent, 76.14 percent, 81.23 percent, 74.12 percent, and 82.19 percent on the Alpha band, Theta band, Beta band, Gamma band, and Delta band using Artificial Neural Networks. To achieve product-specific accuracy, a variety of classifiers have been used.

The highest accuracy on the Delta band is 90.71 percent, followed by 92.21 percent, 82.37 percent, and 83.51 percent using Decision Tree, K-nearest Neighbors, Logistic Regression, and Support Vector Machine (SVM), as shown in Fig. 12. In our study, SVM and ANN performed better than in the previous study [11], and the results obtained are good enough to be used for practical business models.

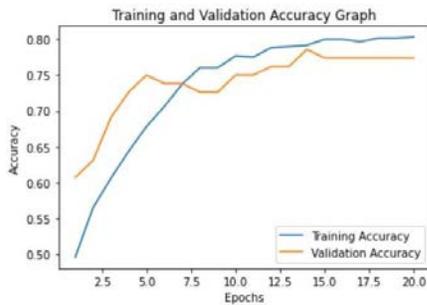


Fig. 11. Training and Validation Learning Curve.

TABLE III. PRODUCT-WISE ACCURACIES ON DIFFERENT BANDS USING ANN

Band	Train Accuracy	Test Accuracy
Alpha	79.73%	78.43%
Theta	71.42%	76.14%
Beta	80.01%	81.23%
Gamma	76.02%	74.12%
Delta	75.17%	82.19%

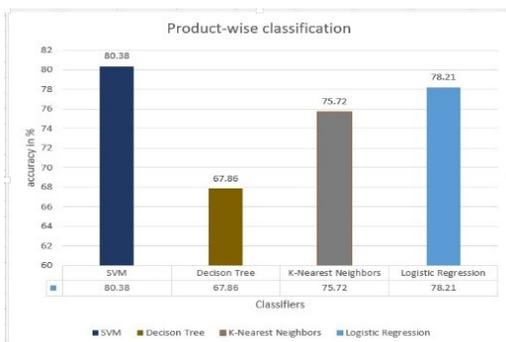


Fig. 12. Accuracies obtained using different Machine Learning Classifiers (Average Accuracy of Products).

D. HEAT Map

The proposed physiological heat map tool allows for the representation of the relative distribution of physiologically inferred emotional or cognitive states of users on a given interface. To make a heat map in MATLAB-based EEGLAB, first select the channel location, then perform the independent component analysis (ICA), and finally plot a 3D component map. Fig. 13 illustrates the 3D heat maps for a consumer's choices. This figure clearly depicts the difference in the heat maps for products with "like" and "dislike". The EEG signals for like are mainly concentrated on the right hemisphere while that for dislike are concentrated on the left hemisphere of the brain.

E. ICA (Independent Component Analysis) Component

The independent components analysis generates a set of weights for all electrodes such that each component is a weighted sum of operation at all electrodes, and the weights are designed to isolate brain electrical signal sources. Components with blink artifacts are possibly the easiest to detect. We have taken a careful approach and only deleted components from the data if you were confident they contained artifacts or noise with no or very little signal. ICA can be used to clean data, separate objects, and exclude certain components from the data, or reduce data. When the individual component analysis is used as a preprocessing method, components can be judged as containing objects based on their topographies, time courses, and frequency spectra. ICA also helps in removing high frequency noise from the EEG signal.

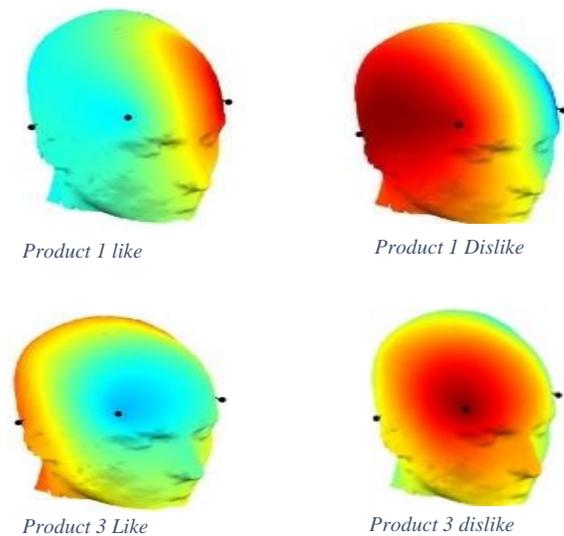


Fig. 13. Heat Map of Consumer Feelings towards the Product.

F. Discussion

In this work, we used EEG data to predict users' product preference using neuroscience. The outcome demonstrates the efficacy of the proposed framework and offers an additional option to existing methods of predicting product market success. This study investigates and improves the classification accuracies of subject-wise and product wise choice preferences. From results it is evident that our proposed

system gives classification accuracy of up to 92.21 percent on the Delta band. The classification accuracies for all five bands are calculated both subject-wise and product-wise. Because there is more randomness in EEG signals of subjects, product-wise accuracy is higher than subject-wise accuracy because EEG signals of the same products are more similar and accurate. The other strong point is that our neuromarketing tool is simple, as we used four dry electrode sensors that can be easily placed on the forehead.

V. CONCLUSION

Using EEG signals, we predicted a customer's product selection preference. The brain activity of 14 participants was recorded while they were viewing products. The Muse 2 headset, which has four sensors, was used to record EEG signals. Further, the filters were applied to make signals smooth and classified using Artificial Neural Networks and other classifiers like SVM, decision tree, logistic regression, and K-Nearest Neighbors. Using all of the above-mentioned classifiers, we obtained subject-and product-level accuracies. Our obtained results demonstrate the effectiveness of the proposed framework, which has provided a superior solution than traditional methods of predicting product success in the market. By extending existing models, the framework can aid in the development of market strategies, research, and forecasting market success.

In the future, this work can be extended by analyzing fictitious responses to product preferences as compared to neutral responses. To improve prediction results, more powerful features and algorithm combinations could be developed.

REFERENCES

- [1] Kumar, H., & Singh, P., "Neuromarketing: An emerging tool of market research," *International Journal of Engineering and Management Research*, pp. 530-535, 2015.
- [2] Kawasaki, M., & Yamaguchi, Y., "Effects of subjective preference of colors on attention-related occipital theta oscillations," *NeuroImage*, 2012, vol.59, no.6, pp. 808-814.
- [3] Sebastian, V., "Neuromarketing and evaluation of cognitive and emotional responses of consumers to marketing stimuli," *Procedia-Social and Behavioral Sciences*, 2014, vol.27, 753-757.
- [4] Lai, C. Q., Ibrahim, H., Abdullah, M. Z., Abdullah, J. M., Suandi, S. A., & Azman, A., "Literature survey on applications of electroencephalography (EEG)," in *AIP Conference Proceedings*, vol. 2016, No. 1, p. 020070, 2018.
- [5] Kumar, H., & Singh, P., "Neuromarketing: An emerging tool of market research", *International Journal of Engineering and Management Research*, vol. 5, no. 6, pp. 530-535.
- [6] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B., "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Eng.*, 2007, vol. 4, no. 2, R1.
- [7] Solhjo, S., Nasrabadi, A. M., & Golpayegani, M. R. H., "Classification of chaotic signals using HMM classifiers: EEG-based mental task classification," In *2005 13th European Signal Processing Con.*, pp. 1-4, (2005).
- [8] Guo, G., & Elgendi, M., "A new recommender system for 3D e-commerce: An EEG based approach," *Journal of Advanced Management Sc.*, 2013, Vol. 1, no.1, pp. 61 – 65.
- [9] Hazarika, N., Chen, J. Z., Tsoi, A. C., & Sergejew, A., "Classification of EEG signals using the wavelet transform," *Signal Processing*, vol. 59, no.1, pp. 61 – 72, 1997.
- [10] Murugappan, M., Murugappan, S., & Gerard, C., "Wireless EEG signals-based neuromarketing system using Fast Fourier Transform (FFT)," In *2014 IEEE 10th international colloquium on signal processing and its applications*, pp. 25-30, IEEE (2014).
- [11] Yadava, M., Kumar, P., Saini, R., Roy, P. P., & Dogra, D. P., "Analysis of EEG signals and its application to neuromarketing," *Multimedia Tools and App.*, 2017, vol. 76, no.18, pp. 19087-19111.
- [12] Chris T. (2018, October 30). Muse 2 review: The world's best meditation tech just got even better. Retrieved from Mashable: <https://mashable.com/article/muse-2-review>.
- [13] Awal, M. A., Mostafa, S. S., & Ahmad, M., "Performance analysis of Savitzky-Golay smoothing filter using ECG signal," in *J. of Com. and Information Technology*, 2017, vol. 1, no. 02, p. 24.
- [14] Skodras, Athanassios. (2015). *Discrete Wavelet Transform: An Introduction*.
- [15] Evgeniou, Theodoros & Pontil, Massimiliano, *Support Vector Machines: Theory and Applications*. Springer: Berlin. pp. 249-257. doi:10.1007/3-540-44673-7_12.
- [16] Maalouf, Maher., "Logistic regression in data analysis: An overview," *International Journal of Data Analysis Techniques and Strategies*, 2011 vol. 3. pp. 281-299. doi:10.1504/IJDATS.2011.041335.
- [17] Patel, Harsh & Prajapati, Purvi., "Study and Analysis of Decision Tree Based Classification Algorithms," *International Journal of Computer Sciences and Eng.*, 2018, vol. 6, pp. 74-78. 10.26438/ijcse/v6i10.7478.
- [18] Cutler, A., D. Cutler, and J. Stevens., *Random Forests*, vol. 45, pp. 157-176, doi:10.1007/978-1-4419-9326-7_5.
- [19] Grossi, Enzo & Buscema, Massimo, "Introduction to artificial neural networks," *European Journal of Gastroenterology & Hepatology*, 2018, vol. 19, pp. 1046-1054. doi:10.1097/MEG.0b013e3282f198a0.
- [20] A. Ali, T. A. Soomro, F. Memon, M. Y. A. Khan, P. Kumar, M. U. Keerio *et al.*, "EEG Signals Based Choice Classification for Neuromarketing Applications," *A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems*, pp. 371-394, 2022.
- [21] Bashir, F., Ali, A., Soomro, T. A., Marouf, M., Bilal, M., & Chowdhry, B. S., "Electroencephalogram (EEG) Signals for Modern Educational Research," in *Innovative Education Technologies for 21st Century Teaching and Learning*, 2021, pp. 149-171, CRC Press.

Tomato Leaf Disease Detection using Deep Learning Techniques

Nagamani H S¹

Research Scholar, University of Mysore, India
Asst. Prof. Dept. of Computer Science, Smt V H D Central
Institute of Home Science, (MCU), Bangalore, India

Dr. Sarojadevi H²

Prof. Dept. Computer Science and Engineering
Nitte Meenakshi Institute of Technology
VTU, Bangalore, India

Abstract—Plant diseases cause low agricultural productivity. Plant diseases are challenging to control and identify by the majority of farmers. In order to reduce future losses, early disease diagnosis is necessary. This study looks at how to identify tomato plant leaf disease using machine learning techniques, including the Fuzzy Support Vector Machine (Fuzzy-SVM), Convolution Neural Network (CNN), and Region-based Convolution Neural Network (R-CNN). The findings were confirmed using images of tomato leaves with six diseases and healthy samples. Image scaling, color thresholding, flood filling approaches for segmentation, gradient local ternary pattern, and Zernike moments' features are used to train the pictures. R-CNN classifiers are used to classify the illness kind. The classification methods of Fuzzy SVM and CNN are analyzed and compared with R-CNN to determine the most accurate model for plant disease prediction. The R-CNN-based Classifier has the most remarkable accuracy of 96.735 percent compared to the other classification approaches.

Keywords—Fuzzy Support Vector Machine (SVM); Convolution Neural Network (CNN); Region-based Convolution Neural Network (R-CNN); color thresholding; flood filling

I. INTRODUCTION

Tomatoes are a major commercial crop grown all over the world. It is sensitive to various illnesses, which reduces tomato quality and yield while also causing significant economic losses. Tomato grey leaf spot is a common disease that damages and kills the leaves of tomatoes, preventing them from growing and producing fruit. The infection that produces grey leaf spots on tomatoes is brutal to remove. Contact, invasion, latency, and onset are the four phases of infection for the pathogen that causes tomato grey leaf spot. As a result, early preventative and control methods are suitable before a large-scale pandemic. Early disease detection can also aid in reducing pesticide usage and pollution, as well as the quality, safety, and health of tomatoes. Traditional disease detection systems cannot meet large-scale planting demands due to low diagnostic efficiency and fast disease transmission, and plants typically miss the appropriate management time [1, 2].

Manually detecting leaf disease with the naked eye needs a team of professionals and ongoing monitoring. When the farm is large, it is costly. As a result, image processing techniques may be used to automatically detect illnesses in leaves, saving time, money, and effort as compared to traditional methods. The early detection of illnesses in leaves improves crop productivity. Disease-affected leaves may be found at an early

stage using image processing techniques like as segmentation, identification, and classification, and crop yield and quality can be improved. Many farmers lack the resources or understanding on how to contact specialists, which makes it more costly, time-consuming, and inaccurate. In this case, the suggested approach proved to be more advantageous in terms of crop observation. The technique is more accessible and less costly when plant illness is detected using leaf symptoms. It takes less time, effort, and precision to use an automated detection technique. Manually detecting leaf disease with the naked eye needs a team of professionals and ongoing monitoring. When the farm is large, it is costly.

As a result, image processing techniques may be used to automatically detect illnesses in leaves, saving time, money, and effort as compared to traditional methods. The early detection of illnesses in leaves improves crop productivity. Disease-affected leaves may be found at an early stage using image processing techniques like as segmentation, identification, and classification, and crop yield and quality can be improved. Many farmers lack the resources or understanding on how to contact specialists, which makes it more costly, time-consuming, and inaccurate. In this case, the suggested approach proved to be more advantageous in terms of crop observation. The technique is more accessible and less costly when plant illness is detected using leaf symptoms. It takes less time, effort, and precision to use an automated detection technique.

Image processing technology can quickly and accurately diagnose illnesses based on their features. Disease prevention approaches may be applied fast, and efforts to avoid additional illnesses can be performed using this strategy. People used to identify tomato ailments based on their own experiences, but the ability to discern between various diseases is limited, and the process is time-consuming. Machine learning and image processing technologies are fast expanding and more widely employed in various industries, including agriculture. The following are two key contributions of this research: R-CNN framework for classifying leaf diseases ii) comparison of different classifiers. The remaining sections of this work are organized as follows: Section 2 did a background survey. Section 3 is about methodology. In Section 4, we give experimental data and analyze it by comparing the best classifiers that may be employed based on the findings of our previous study. Section 5 summarizes the proposed work of this paper.

II. LITERATURE SURVEY

Traditional plant disease detection approaches based on computer vision technologies are commonly utilized to extract the texture, shape, color, and other features of disease spots. This method has a low identification efficiency because it relies on an extensive expert understanding of agricultural illnesses. Many academics have conducted significant research based on deep learning technology to increase the accuracy of plant disease detection in recent years, thanks to the fast growth of artificial intelligence technology [8]. The majority of existing techniques to plant disease analysis are based on disease classification [12].

Mohanty et al. [3] utilized GoogleNet and AlexNet to classify 54,306 plant leaf pictures as healthy or sick in the Plant Village dataset, revealing that GoogleNet had a slightly more significant average classification impact than AlexNet. The trained deep convolutional neural network model has a 99.35 percent accuracy on the test set. Building a deep learning model on a growing and publicly available photo dataset is a simple way to employ clever mobile phones to diagnosis plant diseases in horticulture crops.

Picon et al. [4] employed a deep residual neural network-based upgraded algorithm to identify many plant illnesses in real-world acquisition conditions. For early illness identification, several improvements have been recommended. According to the data, all of the illnesses evaluated had an AuC score of higher than 0.80.

Selvaraj et al. [5, 9] utilized the transfer learning approach to retraining three CNN architectures. Deep transfer learning was utilized to build networks using pre-trained sickness detection models to provide accurate predictions.

Deep learning was proposed by Fuentes et al. [6, 7] for identifying diseases and pests in tomato plant photos acquired at varying camera resolutions. Deep learning meta-architectures, as well as multiple CNN object detectors, were utilized. Data expansion and local and global class annotation were utilized to boost training accuracy and decrease false positives. A large-scale tomato disease dataset was used for end-to-end training and testing. The system correctly detected nine different pests and diseases from the complicated scenarios.

III. METHODOLOGY

Our primary objective is to develop a model to categorize input plant leaf pictures as healthy or unhealthy. The disease kind is also determined if a disease is detected on a plant leaf. Our study compares the R-CNN Classifier to previously established tomato leaf disease detection utilizing fuzzy SVM [15] and CNN [16] Classifiers to detect and categorize tomato leaves suffering from common illnesses. Fig. 1 shows the architecture of the R-CNN-based plant disease detection system. The proposed technique includes image capturing, pre-processing, segmentation, feature extraction, classification, and performance assessment.

A. Dataset Description

The dataset utilized for this investigation has seven primary classifications. Six leaves classes represent unhealthy, while

one represents the healthy leaf class. Each class has 105 examples for a total of 735 leaf images. A classification strategy is required to categorize input photos into one of the classes specified in Fig. 2 for a given image of an apple leaf.

B. Image Pre-processing

The visual noise of the tomato leaf is made up of dewdrops, dust, and insect feces on the plants. The input RGB image is transformed to a grayscale image for accurate results to remedy these concerns. The image size in this circumstance is relatively large, needing image resize. The image size is reduced to 256 * 256 pixels.

C. Image Segmentation

Plant disease detection and categorization rely heavily on image segmentation. The image is simply divided into various things or sections. It analyses visual data to extract information that may be used for further processing. Our prior work [15] is used to accomplish color thresholding and flood filling segmentations.

D. Classification using R-CNN

Rectangular regions are combined with convolutional neural network characteristics in R-CNN (Regions with Convolutional Neural Networks). The R-CNN algorithm employs a two-stage detection procedure. The first stage identifies a set of picture areas that includes a diseased part. In the second stage, each region's object is categorized.

1) *R-CNN procedure*: The following three approaches are employed to build an R-CNN based algorithm, as shown in Fig. 3.

a) To find regions in a photograph that could contain a diseased part. Region suggestions are the names given to these locations.

b) Extract CNN characteristics from the region suggestions.

c) To categorize the objects, use the characteristics that were retrieved.

The R-CNN generates region recommendations using a mechanism similar to Edge Boxes [10]. The proposed elements have been chopped and scaled out of the image. CNN then classifies the clipped and resized regions. Finally, a support vector machine (SVM) trained on CNN features refine the region proposal bounding boxes. A visual illustration of the problem is shown in Fig. 3.

A pre-trained convolution neural network is used to build an R-CNN detector, also known as transfer learning (CNN).

As a starting point for learning a new task, we will use a pre-trained image classification network that has already learned to extract robust and informative features from raw photographs. A portion of the ImageNet database [10], which is used in the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [11], is used in the great majority of pre-trained networks. These networks have been trained on over a million photographs and can categorize a large number of them. Transfer learning with a pre-trained network is typically much faster and easier than training a network from the ground up.

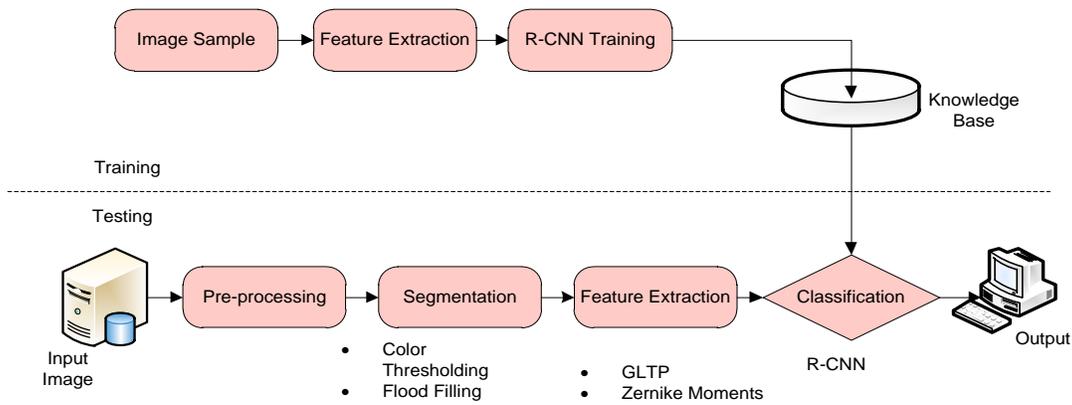


Fig. 1. Architecture of R-CNN-based Plant Leaf Disease Detection.

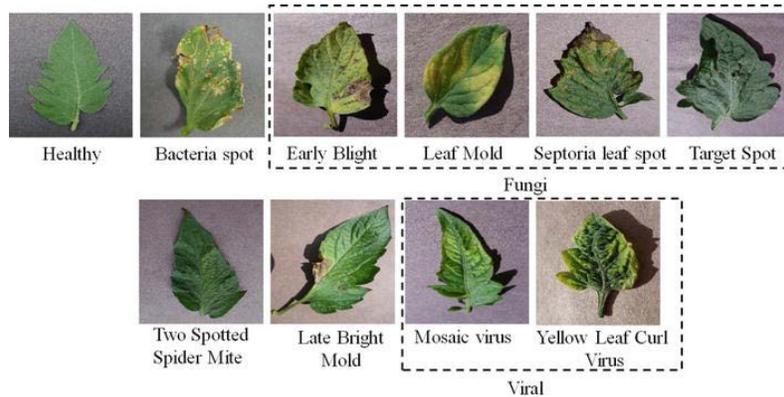


Fig. 2. Tomato Leaf Images with its Diseases.

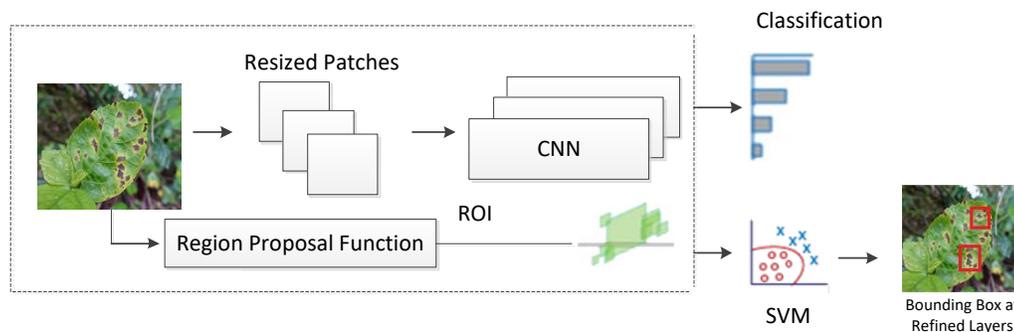


Fig. 3. Process of R-CNN Model.

2) *Transfer learning methodology*: The transfer learning method is employed a pre-built and pre-trained model rather than developing and training a CNN from scratch [13]. Transfer learning is a method based on transferring a model that has been trained on an extensive dataset to a smaller dataset. Early convolutional network layers are retained for object recognition and train the final few levels to generate predictions. The theory is that the convolutional layers extract broad, low-level properties from numerous pictures, such as edges, patterns, and gradients. On the other hand, the last layers identify individual characteristics inside a picture. Fig. 4 depicts the principle of transfer learning [14].

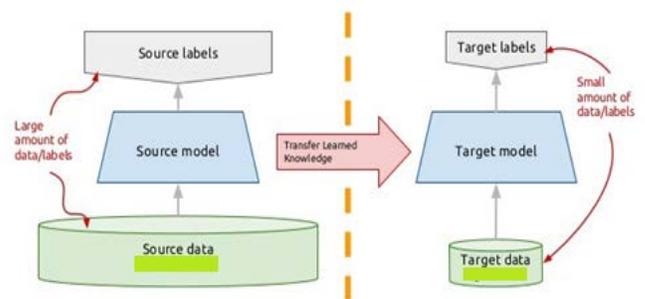


Fig. 4. Concept of Transfer Learning.

A general framework for object recognition transfer learning is as follows:

- a) Insert a CNN model trained on an extensive dataset and is ready to use.
- b) Freeze the parameters of the model's bottom convolutional layers (weights).
- c) Add a custom classifier with many layers of trainable parameters to the model.
- d) The task's training data should be used to train layers of classifiers.
- e) Fine-tune hyperparameters and unfreeze more layers as needed.

This strategy has been proven to be effective in several sectors. It is a terrific tool to have in our arsenal, and it is generally the first thing we do when confronted with a new image recognition challenge.

3) *Transfer Learning Characteristics:*

- a) It requires transferring information from one source task to the target task to learn and grow.
- b) DNN trained on raw pictures has been observed to demonstrate an unusual occurrence in which the network's initial layer appears to gain Gabor filter-like properties.
- c) The first layer's characteristics were found in a variety of datasets.
- d) The initial layers' general characteristics disregard the picture dataset, task, and loss function.
- e) The filters learned by specific ResNet layers can be reused by other ResNet layers if they learn the same feature. In convents, this practice is called extremely successful transfer learning.
- f) While overcoming data shortage, transfer learning saves time and money in training.

4) *R-CNN model design:* To generate R-CNN models based on a previously trained CNN for image classification. The R-CNN model is built on the foundation of a pre-trained network. The last three categorization levels are eliminated, and new layers customized to the item types wish to detect are added in their place.

As shown in Fig. 5, the final three layers in this network are fc1000, fc1000 softmax, and ClassificationLayer fc1000; the last three levels in this network are fc1000, fc1000 softmax, and ClassificationLayer fc1000. Fig. 6 depicts the removal of the final three layers.

As indicated in Fig. 7, add the new categorization layers to the network. The layers are put up to categorize the number of items that the network should detect.

5) *Convolution neural layers of RCNN model:* The input layer, the initial layer of the R-CNN architecture, delivers raw data to the network. The 32*32*3 raw image was recorded. The properties are extracted via the convolution layer, the transformation layer. As a result, this layer's inputs were leaf pictures. The R-CNN model's first convolution layer consists of 32 different 3*3 filters, each having one stride and one

padding. After the convolution layers, the rectified linear unit layer (ReLU) is the most often utilized rectifier unit for neuron outputs. Pooling, commonly employed after the ReLU layer, decreases the subsequent convolution layer's input size (width and height). Fig. 8 depicts the entire structure.

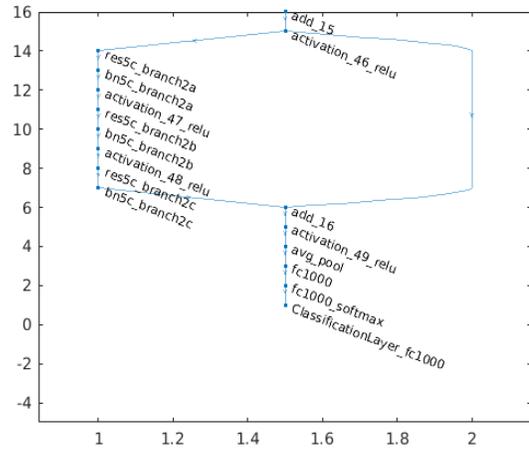


Fig. 5. Layers of ResNet-50 Pre-trained Network.

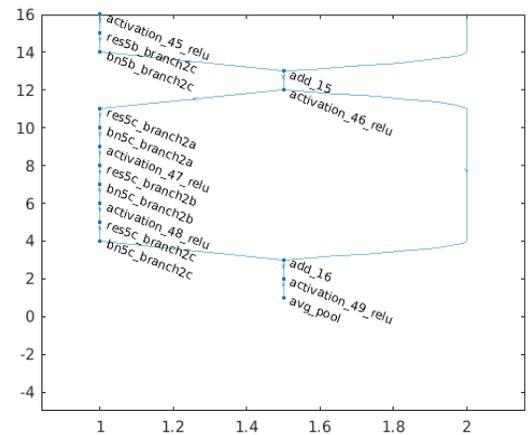


Fig. 6. Layers of ResNet-50 Network by removing Three Layers.

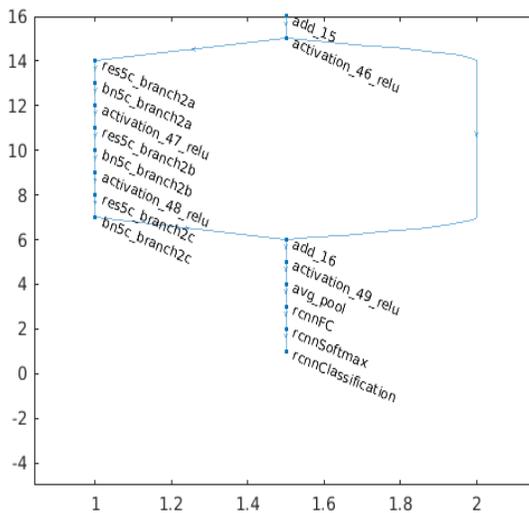


Fig. 7. Layers of ResNet-50 Network after addition of New Layers.

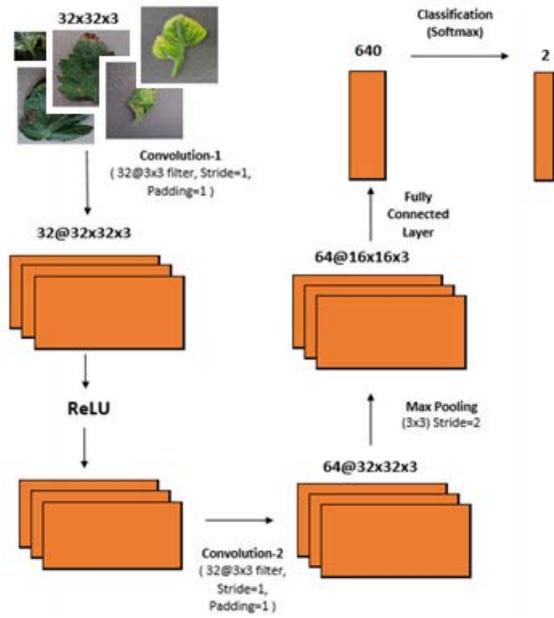


Fig. 8. Architecture of CNN Inside R-CNN.

The sub-maximum area's value was forwarded to the max-pooling layer. The scale value must be carefully controlled to avoid the attribute size reducing too quickly, the image attributes becoming rough, and the attributes getting lost in the max-pooling layer. There are layers for convolution, ReLU, and max-pooling before the finally linked layer. A 64-node fully connected layer was employed in the R-CNN. As a result, this layer is linked to the prior levels' core constituents. This layer is followed by the classification layer, which performs classification. A softmax classifier is a multi-class classifier based on an extension of the logistic function. The output map is connected to the Softmax function, the final layer of the proposed architecture, through the final convolution layer. It establishes the likelihood of each class.

IV. EXPERIMENTAL RESULT AND ANALYSIS

This section evaluates the performance of several classification techniques such as Fuzzy SVM, CNN, and R-CNN and shows that the Fuzzy SVM classification methodology surpasses the others. The datasets for tomato leaf disease are split into two categories: training data (70%) and testing data (30%). The Tomato Plant Disease dataset contains 735 pictures and seven classes named Bacterial Spot, Mosaic Virus, Yellow Leaf Cur Virus, Early Blight, Late Blight, Leaf Mold, and Healthy.

The implementation of these classifiers takes place in MATLAB 2019B. The input images as illustrated in Fig. 9, are pre-processed in this experiment to decrease noise, and then segmentation is performed using Color-thresholding and flood filling. The contaminated section of the leaf is then removed, and the GLTP and Zernike moments of that infected area are determined. Then, Classifiers are employed to determine the illness name.

The confusion matrix is a metric for evaluating the performance of a machine learning classification task with two or more classes as output. There are four distinct combinations

of projected and actual values in this table. It's great for assessing things like recall, precision, specificity, accuracy, and, most crucially, AUC-ROC curves. The confusion matrix of R-CNN is shown in Fig. 10.

A Receiver Operator Characteristic (ROC) curve is a graphical representation of a classifier's diagnostic capabilities. The ROC curve depicts the sensitivity vs. specificity trade-off. Classifiers with curves that are closer to the top-left corner perform better. The ROC curve that was developed for R-CNN is shown in Fig. 11.



Fig. 9. RCNN-Trained ResNet 50 Sample Results.

R-CNN Confusion Matrix

	01	02	03	04	05	06	07	
01	105							100.0%
02		103	1		1			98.1% 1.9%
03			105					100.0%
04				104	1			99.0% 1.0%
05					104	1		99.0% 1.0%
06		1				104		99.0% 1.0%
07		1					104	99.0% 1.0%
	100.0%	98.1%	100.0%	99.0%	100.0%	97.2%	100.0%	
		1.9%		1.0%		2.8%		
	01	02	03	04	05	06	07	
	Predicted Class							

Fig. 10. Confusion Matrix of R-CNN Classifier.

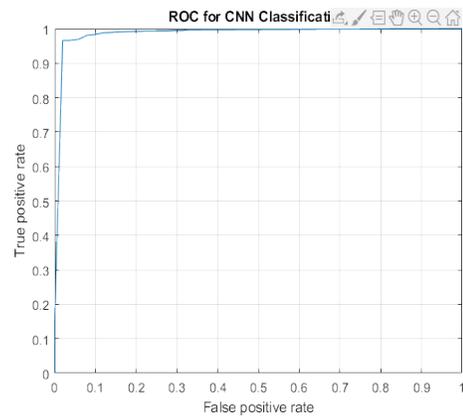


Fig. 11. ROC Curves for R-CNN Classifier.

The confusion matrix of CNN, and fuzzy-SVM, respectively, are shown in Fig. 12, and 13.

Fig. 14 depicts the classification accuracy for all the 3 classifiers as a comparison analysis. The leaves are classified as healthy or unhealthy and the type of condition if they are diseased.

Precession, recall, and accuracy are the other performance indicators shown in Fig. 15 for all of the classification algorithms for tomato leaf disease datasets.

True Class \ Predicted Class	01	02	03	04	05	06	07	Accuracy	Loss
01	104						1	99.0%	1.0%
02		102		1	2			97.1%	2.9%
03	1	1	98			3	2	93.3%	6.7%
04				104		1		99.0%	1.0%
05		5	1		90	5	4	85.7%	14.3%
06		1		1	5	96	2	91.4%	8.6%
07		3	5		2	3	92	87.6%	12.4%
	99.0%	91.1%	94.2%	98.1%	90.9%	88.9%	91.1%		
	1.0%	8.9%	5.8%	1.9%	9.1%	11.1%	8.9%		

Fig. 12. Confusion Matrix of CNN Classifier.

True Class \ Predicted Class	1	2	3	4	5	6	7	Accuracy	Loss
1	88	1	3		2	5	6	83.8%	16.2%
2		97		1	5	2		92.4%	7.6%
3	5	13	66			4	17	62.9%	37.1%
4				96	2	4	3	91.4%	8.6%
5	12	30	1	3	45	6	8	42.9%	57.1%
6	5	17	8	8	22	35	10	33.3%	66.7%
7	12	13	3		7	5	65	61.9%	38.1%
	72.1%	56.7%	81.5%	88.9%	54.2%	57.4%	59.6%		
	27.9%	43.3%	18.5%	11.1%	45.8%	42.6%	40.4%		

Fig. 13. Confusion Matrix of Fuzzy SVM Classifier.

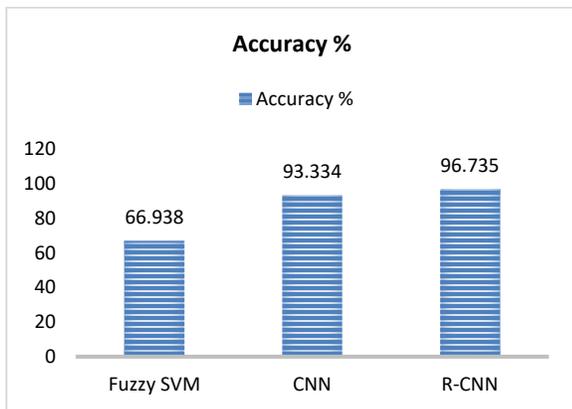


Fig. 14. Comparison between Classification Techniques.

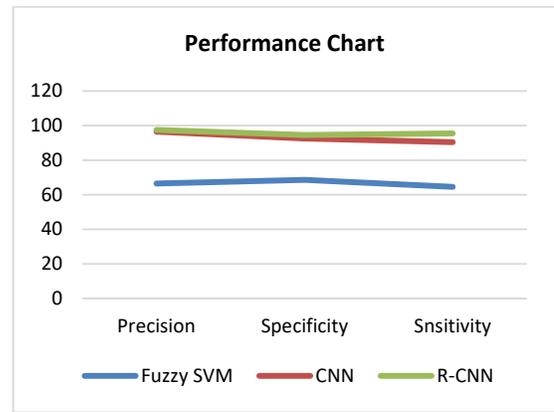


Fig. 15. Precision, Specificity and Sensitivity of Classifiers.

Previously developed models are used for performance comparison using the confusion matrix. The following is a quick summary of these models: The confusion matrix for R-CNN findings for 735 input leaf samples is shown in Fig. 10.

A healthy and unwell people database is produced in a Fuzzy SVM-based Classifier [15]. There are six distinct illnesses and 735 photos in the classes in the input database. Color thresholding and flood filling are the two segmentation techniques used. For greater accuracy, the output of both algorithms is integrated using ROI (region of interest). The diseased section of the leaf's characteristics was extracted using the segmented picture. The Color Covariance Vector (CCV) and Gradient Direction Pattern (GDP) algorithms are employed. This phase extracts a total of 56 color features and 256 gradient characteristics. With 420 photos in four classes and a parentage accuracy of 97.142, this Fuzzy SVM algorithm provides the best results. However, increasing the number of photos from 420 to 735 reduces performance to 66.938 percent. A CNN-based classifier is employed to increase performance.

The CNN-based Classifier [16] includes 735 healthy and ill plant leaves. Among the seven forms of tomato leaf diseases in the database are Bacterial Spot, Mosaic Virus, Yellow Leaf Cur Virus, Early Blight, Late Blight, and Leaf Mold. The images were all shot in a lab environment. The images of leaves were divided into two categories: training and testing. Leaf images are separated into two groups to test performance: 70–30 (70 percent of the photos are for training, and 30 percent are for testing). The segmented section is used to retrieve the unhealthy portion of the leaf.

V. CONCLUSION

Machine learning and image processing technology benefits from traditional manual diagnostic and recognition procedures for crop disease diagnosis. A small number of unhealthy image samples is necessary. Deep learning is one of the most well-known methods of artificial intelligence. Computer vision is one of the many fields where deep learning is used. It is capable of picture categorization, object identification, and semantic segmentation.

The disease diagnostic approach based on the deep convolutional network (CNN) minimizes comprehensive picture pre-processing and feature extraction methods

compared to standard pattern detection techniques. It employs an end-to-end structure to simplify the detecting procedure. In this study, we compared the best classifier for effectively classifying tomato leaf disease with an accuracy of 96.735 percent.

VI. CONFLICTS OF INTEREST

The authors declare no conflict of interest.

REFERENCES

- [1] Mariko T, Hiroshi E. How and why does tomato accumulate a large amount of GABA in the fruit? *Front Plant Sci.* 2015; 6:612.
- [2] Fuentes A, Yoon S, Youngki H, Lee Y, Park DS. Characteristics of tomato plant diseases—a study for tomato plant disease identification. *Proc Int Symp Inf Technol Converg.* 2016; 1:226–31.
- [3] Mohanty SP, Hughes DP, Salathé M. Using deep learning for image-based plant disease detection. *Front Plant Sci.* 2016; 7:1419.
- [4] To EC, Li Y, Njuki S. A comparative study of fine-tuning deep learning models for plant disease identification. *Comput Electron Agric.* 2018; 161:272–9.
- [5] Picon A, Alvarez-Gila A, Seitz M, Ortiz-Barredo A, Echazarra J, Johannes A. Deep convolutional neural networks for mobile capture device-based crop disease classification in the wild. *Comput Electron Agric.* 2019;1(161):280–90.
- [6] Selvaraj MG, Vergara A, Ruiz H, et al. AI-powered banana diseases and pest detection. *Plant Methods.* 2019; 15:92.
- [7] Zhong Yong, Zhao Ming. Research on deep learning in apple leaf disease recognition. *Comput Electron Agric.* 2020; 168:105146.
- [8] Aravind KR, Raja P, Anirudh R. Tomato crop disease classification Using A Pre-Trained Deep Learning Algorithm, *Procedia Comput Sci.* 2018; 133:1040–7.
- [9] Karthik R, Hariharan M, Anand Sundar, Mathikshara Priyanka, Johnson Annie, Menaka R. Attention embedded residual CNN for disease detection in tomato leaves. *Applied Soft Comput.* 2020.
- [10] Girshick, R., J. Donahue, T. Darrell, and J. Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." *CVPR '14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition.* Pages 580-587. 2014.
- [11] Fuentes A, Yoon S, Kim SC, Park DS. A robust deep-learning-based detector for real-time tomato plant diseases and pest's recognition. *Sensors.* 2022; 2017:17.
- [12] D. T. Mane and U. V. Kulkarni, "A survey on supervised convolutional neural network and its major applications," *International Journal of Rough Sets and Data Analysis*, vol. 4, no. 3, pp. 71–82, 2017.
- [13] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017*; pp. 4700–4708.
- [14] Wang, J., Chen, L., Zhang, J., Yuan, Y., Li, M., Zeng, W., 2018. In *Chinese Conference on Image and Graphics Technologies*, Springer, Cnn transfer learning for automatic image-based classification of crop disease. pp. 319–329.
- [15] Nagamani H S and Dr. Sarojadevi H "Leaf Diseases Detection using Fuzzy Classifiers" Unpublished.
- [16] Nagamani H S and Dr. Sarojadevi H "Leaf Disease Identification by Extracting Gradient Local Ternary Pattern & Zernike Moment" Unpublished.

Feature Selection Pipeline based on Hybrid Optimization Approach with Aggregated Medical Data

Palwinder Kaur¹

Department of Computer Science
IKG Punjab Technical University, Jalandhar, Punjab, India

Rajesh Kumar Singh²

Department of Computer Applications
SUS Institute, Tangori, Punjab, India

Abstract—For quite some time, the usage of many sources of data (data fusion) and the aggregation of that data have been underappreciated. For the purposes of this study, trials using several medical datasets were conducted, with the results serving as a single aggregated source for identifying eye illnesses. It is proposed in this paper that a diagnostic system that can detect diabetic retinopathy, glaucoma, and cataract can be built as an alternative to current methods. The data fusion and data aggregation techniques used to create this multi-model system made it conceivable. As the name implies, it is a way of compiling data from a large number of legitimate sources. The development of a pipeline of algorithms was accomplished through iterative trials and hyper parameter tweaking. CLAHE (Contrast Level Adaptive Histogram Equalization) approaches, which increase the gradient between picture edges, improve segmentation by raising the contrast between picture edges. The Gabor filter has been shown to be the most effective method of selecting features. The Gabor filter was selected using a hybrid optimization method (LION + Cuckoo), which was developed by the author. For automation, the Support Vector Machine (SVM) radial is the most effective method since it delivers excellent stability and accuracy in terms of accuracy and recall, as well as precision and recall. The discoveries and approaches detailed here provide a more solid foundation for future image-based diagnostics researchers to build on in the future. Eventually, the findings of this study will help to improve healthcare workflows and practices.

Keywords—Content-based image retrieval system; CLAHE; Gabor filter; Cuckoo search; LION optimization; support vector machine

I. INTRODUCTION

The covid-19 pandemic has forced rethinking about the effectiveness of clinical workflows and practices [1]. The current advances in image processing and machine learning algorithms make it imperative that new algorithms and methods be incorporated into health care technologies and systems[2]. Current research trends and evidence point out that creating multi-model systems requires using multiple protocols, stacks of technologies, and an array of algorithms and multiple data sources[3]. Choosing a specific technology stack and pool of algorithms for building a reliable system has become tedious work and confusing. The primary reason is the availability and choice of ready-to-use frameworks, APIs, libraries, and technological stacks. Hence, finding and appropriating existing algorithms requires exhaustive

experimentation and optimization e.g., at the data processing level: importing, validating, cleaning, converting, normalizing, and pre-processing the data requires a lot of experience and intuitiveness for selecting the suitable method, which would yield the best possible outcomes [4]. Application of methods such as data fusion, aggregation and argumentation also need to be explored, especially when the number of data instances of particular class are less and there is an imbalance in the dataset.

The term "pipeline" in computer science has broader connotations in the current context [5]. It is referred to as multi instructions performed as a unit. The computer unit may be some software module or hardware such as Graphics Cards. HTTP pipeline is a sequence of steps taken to handle HTTP traffic and tasks in the context of the research work. "Algorithm Pipelining" is more appropriate as this research work involves constructing algorithms, frameworks, and systems sequences that can perform tasks such as prediction of corona virus with the help of experiments with high reliability [6]. A pipeline is another method of defining an experiment [7]. An experiment whose objectives are known and defined and the outcome helps construct a fully functional system. In this research work, an attempt will be made to identify an accurate workflow of the methods (image processing and machine learning) that would yield a high-performing system that can detect at least three types of eye diseases, i.e.; diabetic retinopathy, glaucoma, and cataract. Hence, in the next section, the workflows, approaches, and methods are discussed which are used these days to construct multiple-model eye disease detection systems.

II. REVIEW

Medical imaging technology has significantly progressed, which has helped to reduce the burden of detection of numerous diseases. Machine learning algorithms and their comprehensive frameworks have greatly helped the image segmentation field [8][9]. However, the biggest problem that the researcher faces in this context is building diagnostic systems related to the characteristics of the data set [10]. By analyses of the publically available medical image data set e.g. eye diseases, it can be observed that many of these data sets belong to particular or specific modalities, and at the same time, they are poorly annotated [11][12]. Many data sets are not labelled as per the stage of the disease; in many cases, the

dataset is not as per the requirement of machine learning modelling [13]. Due to this challenge, multi-model disease detection systems are hard to realize [14]. In simple words, it means that highly specialised systems of detection can be constructed. However, detection system for a specific domain is hard to construct; for example, the current literature quotes many examples of handling diabetic retinopathy detection systems[15], but few new systems are illustrated in high impact journals that deal with the detection of multiple diseases such as glaucoma, cataract, and diabetic retinopathy at the same time [16][17][18]. This industry faces two kinds of problems: the first is limited annotated data sets, and the second one is weak or incomplete annotations in the data set as per medical grade system. Contemporary literature cites different solutions to overcome the problem depending upon the problem.

The most frequently used technique is data aggregation [19], data augmentation [20][21] and data fusion [22]. In data augmentation, the existing data set size is increased by adding more synthetic data to it, or learning from the existing annotated data is done for constructing a more significant size data set; this way, multiple diseases and modalities can be covered. Such methods also help overcome the imbalance in the data set and help leverage active learning. The most significant advantage is that multi-disease detection systems can now be constructed using multiple data sets or data fusion techniques. Data fusion algorithms leverage multiple disease data sets for constructing image processing functions that work on heterogeneous disjoint sets that can support multi-disease detection systems [23]. Some authors refer to such procedures as data adaptation also. Data adaptation is a process by which a data set is constructed, which helps the learning component of the detection system to discriminate between the various disease modalities and come out with an effective solution. Data augmentation, data fusion, and data adaptation help immensely overcome the challenge of building generic systems of detection [24]. However, the current literature also points out that there are limitations in using single algorithms for building multi-disease detection systems. Research in this context also shows that building classification systems relying on specific features and single classification methods may not yield a stable numerical system. There is always a need to use various methods and solutions for detecting multiple diseases. Hence, many researchers have concluded that the usage of hybrid techniques and combination approaches is far better than training a specific machine learning model. This way, a robust model for constructing multiple disease detection systems can be realised and implemented.

From the current literature it is amply clear that as a strategy for building medical detection systems, three possible path ways can be used for constructing systems of disease detection [25]. The first uses purely statistical methods, the second uses optimization methods, and the third uses machine learning algorithms or deep learning models. It should be however be noted that Image segmentation is a precursor for using these three approaches because extracting the object of interest from the medical images is a fundamental step in

building disease systems. An important gap that is generally visible in the current literature is that few scholars are building systems can multiple diseases detection in medical domain. Generally, the focus of research paper is to work with a single medical modality with specific dataset. However, the need to hour is to construct models that can automatically detect multiple ailments in a comprehensive way. It became critical in context of detecting eye problems due to covid-19 pandemic norms.

In short, it can also be observed from the current research works in context of most relevant approaches are statistics, machine and deep models. Statistical methods such as descriptive statistics, correlation, f-test, t-test, etc., are generally used to understand the nature of the data and identify the suitability of the data for machine learning model [26]. The optimisation algorithms [27][28] such as Genetic Algorithm, ant-colony-optimisation [29], differential-evolution,cuckoo-search [30], particle-swarm-optimisation, firefly, metaheuristic swarm-optimisation, Harris-hawks-optimisation, bat-algorithm, lion-optimiser, grey-wolf-optimiser, moth-flame-optimisation, flower-pollination-algorithm whale-optimisation-algorithm, etc. are used for constructing feature engineering hypothesis for attaining the best possible solutions [31][32][33]. The automation of diagnose comes with implementing machine learning and deep learning algorithms. Current literature gives ample evidence that authors are primarily citing hybrid methods for producing high-precision systems of medical disease detection [34][35]. Large amounts of citations can be found that are showing the most frequently used machine learning algorithms for detecting eye problems include K-Means, K-nearest neighbour (KNN), support vector machine (SVM), ANN or neural networks , decision trees, logistic regression.

This research attempts to analyse three pipelines that would yield a numerically stable multi-disease detection system. The selection of methods used in each type of pipeline is based on the previous research works done by contemporary technical people. Secondly, it is a sincere attempt to find a novel pipeline of methods that can offer consistently repeatable performance detecting multiple eye diseases.

III. MATERIALS AND METHODS

In this section, the steps that make up the workflow of this research are given. It explains techniques, procedures, and algorithms used for building a system designed for eye diagnostics. The research flow block diagram Fig. 1 may be referred to for better understanding. The dataset used in this research work is publically available (https://github.com/palavibhangu/retina_dataset.)The dataset has 300 images of each of three types (diabetic retinopathy, cataract, glaucoma) of eye diseases and healthy eye images.

The aggregated size of the dataset of 1200 images was realized with the help of operation referred as data fusion. As mentioned earlier, in data fusion, multiple datasets are stacked and organised to act as single source of dataset. This has been done to overcome the challenge of low availability of particular class of instances of medical data.

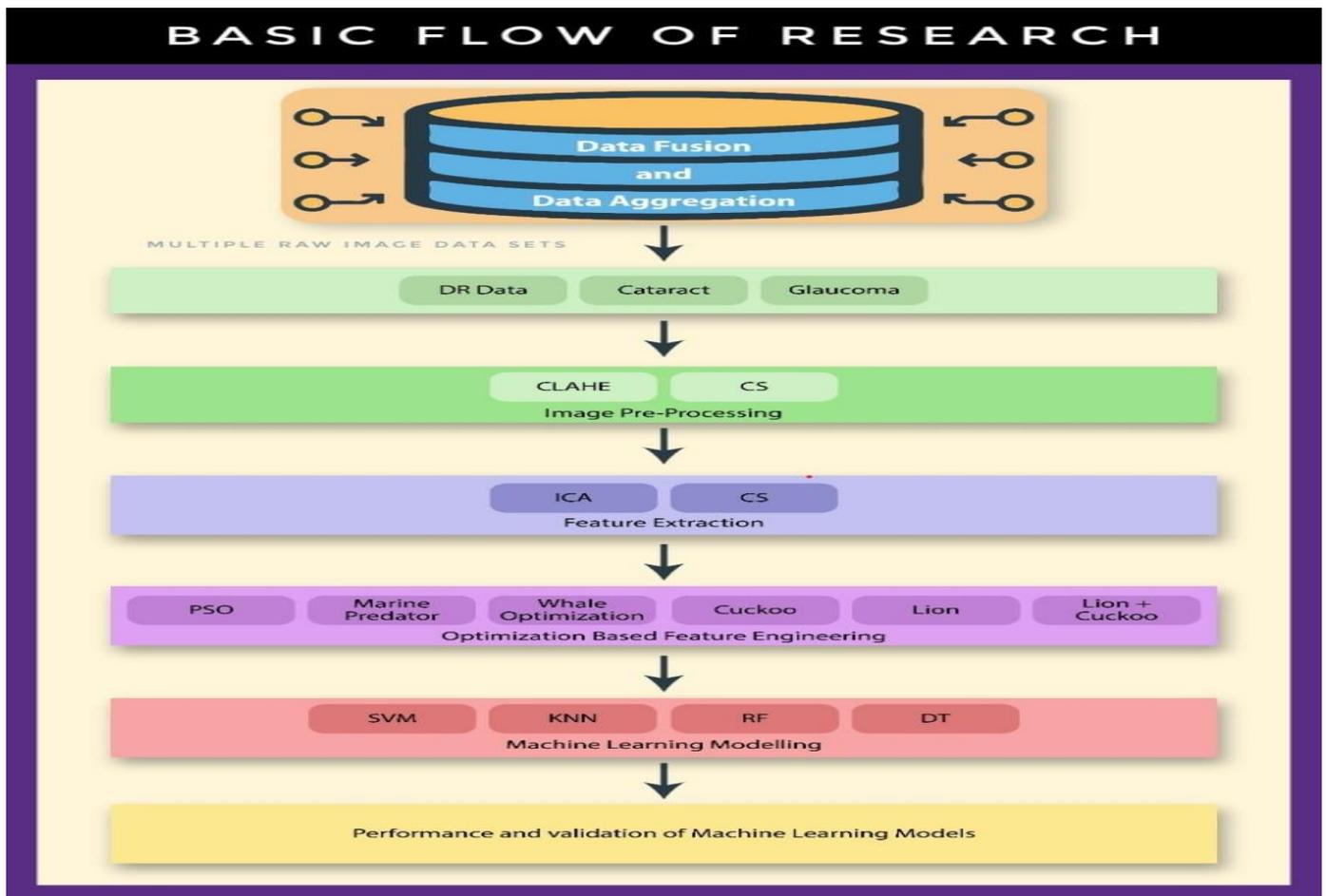


Fig. 1. Basic Flow of Research.

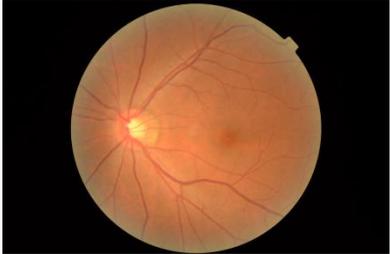
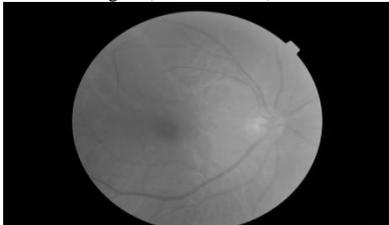
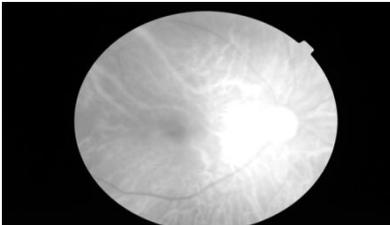
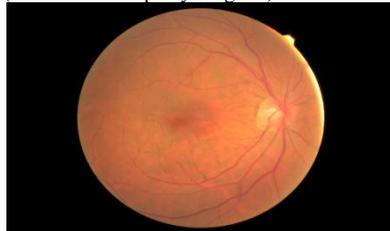
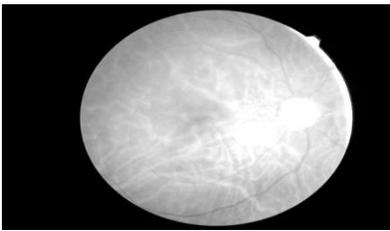
A. Image Processing

In this section, demonstration of the pre-processing steps that include selection of appropriate contrast methods and performance assessment of this step is discussed. It is apparent that for building a generic system of eye problem detection, the images were subjected to pre-processing operations such as contrast enhancement. The purpose of the contrast enhancement is to increase the differential between the various segments of the images. Increasing the differential between the object's pixels that have higher values will attain higher intensity levels and the similarity of the lower-intensity pixels will acquire lower levels of intensity. This process is quite helpful when the segmentation process has to be done as an essential step. Therefore, technically it refers to any technique that uses a function to exaggerate the apparent difference between adjacent structures created during image processing. This helps characterize objects of interest, i.e., a characteristic that can hint at eye problems. Multiple algorithms are available in the image processing domain to improve the images' medical quality. These include histogram-based methods such as adaptive histogram equalisation and contrast

stretching (CS) methods such as min-max stretching. In the context of the problem undertaken, after a lot of experimentation and quality grading, it was found that the CLAHE method[36] is most suitable for the said purposes. The CLAHE method has step operations with which it increases the contrast. The first step divides the image into tiny regions and creates a local histogram for each region. A map of the local histogram is constructed. After this, a clipping point of the histogram is identified for each region. As the contrast process iterates, every region's noise is also reduced with the help of the subtraction method. The result is redistribution of the intensity values of the image. Table I gives the output of the CLAHE.

It can be observed from the comparison that the CLAHE algorithm produces better levels of differentiation between the various entities embodied in the fundus eye images. The selection of the CLAHE algorithm is based on the assessment of output given in the next section. It should also be noted that classical histogram equalization method was initially evaluated and it was found that dynamic or adaptive methods always perform better.

TABLE I. CONTRAST OUTPUT TABLE

Original Image(s)	CLAHE	1) <i>Min-Max Contrast Stretching</i>
(Healthy Original) 	Image I (Healthy 012) 	
(Glaucoma Original) 	CLAHE Image I (Glaucoma 01) 	
(Cataract Original) 	CLAHE Image I (Cataract 007) 	
(Diabetic Retinopathy Original) 	CLAHE Image I (Diabetic Retinopathy 082) 	

It is difficult to objectively evaluate the quality of contrast that an algorithm may provide in photographs from a technical standpoint. The use of a subjective judgment of the image is more appropriate in these situations. The advantage is that domain experts will make decisions in accordance with the medical grade standard of care. Thus, two judges (Judge 1 & Judge 2) were assigned the task of evaluating four factors related to the quality of the images changed by contrast algorithms: distortion in the image, artefacts, noise and information gain that can be valued after the contrast enhancement operation. A questionnaire was developed, and judges assigned scores between 1 and 3 on a scale of 1 to 3. The number 3 indicates that there is no introduction of noise,

distortion, or artefacts as a result of contrast. The number 1 represents the presence of 100 percent noise, distortion, or artefacts in the freshly produced images. If a number 2 is assigned, the value signifies a 50 percent chance of noise, distortion, and artefacts occurring. The same is true for the factor information gain: one indicates that there is no information gain when the contrast algorithm is performed, and 3 indicates that there is a 100 percent gain in the information, indicating that the contrast transformation will be beneficial in better segmentation. There were two experts participated in the evaluation process, and the inter-rater agreement (using average score) between them was computed, since, there are four medical modalities the results are shown in Tables II, III, IV and V, respectively.

TABLE II. HEALTHY IMAGES (RANDOM SAMPLE = 25)

Healthy Samples CLAHE vs CS				
	Factors	Judge 1 Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.68	2.6	2.64
	Distortion	2.68	2.6	2.64
	Artefacts	2.60	2.68	2.64
	Information Gain	2.84	3	2.92
CS	Noise	2	2	2
	Distortion	2	2	2
	Artefacts	2	2	2
	Information Gain	2	2	2

TABLE III. GLAUCOMA IMAGES

Glaucoma Samples CLAHE vs CS				
	Factors	Judge 1 Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.64	2.62	2.63
	Distortion	2.68	2.64	2.66
	Artefacts	2.92	2.92	2.92
	Information Gain	2.92	2.92	2.92
CS	Noise	2.1	2.1	2.1
	Distortion	2	2	2
	Artefacts	2	2.3	2.1
	Information Gain	2.4	2.4	2.4

TABLE IV. CATARACT IMAGES

Cataract Samples CLAHE vs CS				
	Factors	Judge 1 Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.68	2.64	2.64
	Distortion	2.92	2.68	2.8
	Artefacts	2.8	2.8	2.8
	Information Gain	2.76	2.8	2.78
CS	Noise	2.2	2.2	2.2
	Distortion	2	2	2
	Artefacts	2.2	2	2.1
	Information Gain	2	2.2	2.1

TABLE V. DIABETIC RETINOPATHY IMAGES

Diabetic Retinopathy Samples CLAHE vs CS				
	Factors	Judge 1 Mean Score	Judge 2 Mean Score	Average Score
CLAHE	Noise	2.8	2.68	2.78
	Distortion	2.8	2.8	2.8
	Artefacts	2.76	2.68	2.72
	Information Gain	2.88	2.88	2.88
CS	Noise	2	2	2
	Distortion	2.2	2	2.1
	Artefacts	2	2.4	2.2
	Information Gain	2.1	2.1	2.1

Observations from Table II to V demonstrate that CLAHE method is more effective than contrast stretching. In all healthy images, Glaucoma, Diabetic Retinopathy and Cataract, the evaluation shows the CLAHE method is the most stable and reliable algorithm for the said purpose. This may be attributed to the fact; the correct parameters were selected for

taking maximum advantage of the CLAHE algorithm. The parameters; Windows size = 8, Clip limit =0.4, Bin size 255) of CLAHE and use of Rayleigh (alpha value =0.35) based distribution for construction of the histogram yield a better output. Min-Max Contrast Stretching is intensity normalization; this is a typical well established pre-processing step taken by many researchers; nevertheless, in the current

context, it is performing not well as compared to the CLAHE method. In the case of Min-Max Contrast stretching intensity increases but loss of information/pixels is also happening. It is now time for extracting features from these quality enhanced image dataset. The coming section discuss the process of extracting and selecting appropriate optimization algorithm that produces highest possible accuracy of the detection system that is based on machine learning model.

B. Optimization based Feature Engineering

It is possible to take full advantage of machine learning when one looks for recognisable patterns in large quantities of data. With conventional statistics, data consolidation and reduction is the key, and the quality of diversity of the data is given a lower mark. However, machine learning depends on extensive data and high levels of detail (think variety) (think columns or attributes). *Feature engineering* is used to get manual and automated analyses to speed up by adding more features/attributes and providing more details on existing data[37]. Feature analysis can help developers exploit and investigate data with more profound patterns. They are helpful for many machine learning procedures and vital to spot trends that can give real-time hints on diseases in our context. There are two main ways to expand features: ingesting more data after pre-defined features are created or training data to increase available features. The feature selection process includes selecting combinations of variables with large discriminative values to support the detection of various types of classes in the dataset.

As indicated in recent publically available literature assessments, critical variables for diagnosing eye abnormalities include an examination of the eyes' colour, texture, and form. The Gabor Filter was used to analyse the texture [38][39] and it was compared to a independent component analysis (ICA) method [40] for determining the most acceptable features in the image dataset. ICA assists in the discovery of a reduced projection picture or sub space of the original image with decreased dimensions. This reduces overhead while extracting the best feasible statistically independent information from each image. Additionally, ICA method encompasses a wide range of kurtosis and skewness values. The fixed objective function is determined by the differential equation (1).

$$f'(x) = (x-a) f(x) / (b_0 + b_1 x + b_2 x^2) \quad (1)$$

Where a, b₀, b₁, and b₂ are distribution parameters. When the source distributions are known (as they are in this scenario), the score functions are the ideal choice for the objective function. The Pearson ICA system's scoring function is defined as $f'(x) = -f'(x) / f(x) = (x-a) / (b_0 + b_1 x + b_2 x^2)$. The parameters a, b₀, b₁, and b₂ are estimated using the moments approach.

The Gabor filter helps to extract features of images by computing features at different frequencies and by changing the theta angle. This way features from multiple orientations and directions are extracted. Mathematically, it is computed using equation (2).

$$g(x, y, \lambda, \theta, \phi, \sigma, \gamma) = \exp\left(\frac{-x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) * \cos(2\pi \frac{x'}{\lambda} + \phi) \quad (2)$$

Where

$$x' = x \cos \theta + y \sin \theta, y' = -x \sin \theta + y \cos \theta$$

After applying the Gabor filter and extraction of Gabor vector and ICA components, the dataset was transformed into a feature matrix of 1000x100 with the help of the reshape function of Matlab. This was done so that uniform sets of features are used from each image for machine learning. It is expected that as a result of application of these feature selection methods, the chosen features will have a smaller classification error and a higher degree of generalisation when machine models will be constructed.

The analysis for finding the best optimisation algorithm for feature selection depends on four performance factors. The first one is the coverage percentage: it constantly desired that it should be 100% so that no feasible area during the optimisation process is left uncovered to find the optimal solution to the problem. For example, the PSO is not a global optimisation algorithm [41]; hence, it cannot guarantee convergence to a local optimum. Due to this fact, the stability of the solution may be questionable. The excellent level of coverage is reflected in the accuracy parameter. The second performance analysis is about the computational time the optimisation solution takes to reach the most feasible solution. It is better to have feasible solutions fast. Then, the number of features that the algorithm finds useful is critical. It is generally expected that the lower the number of features, the lower is the overhead for the machine learning algorithms for building an appropriate solution.

All these optimization algorithms were executed using standard parameters values such as the number of iterations = 10 and solutions=100. Each algorithm has specific parameters that need to be configured before these algorithms can be executed. It can be observed from Table VI that cuckoo and LION algorithms [42] are most competent in terms of accuracy and number of selected features. However, it is better to use a hybrid approach and combination of the LION and cuckoo as it further reduces the overhead and keeps the accuracy levels a bit higher than individually using the LION or Cuckoo algorithm. The optimization algorithms' performance analysis shows that applying a hybrid algorithm helped obtain the best possible solution in terms of the number of features. The coverage of the hybrid algorithm is excellent, which lead to the selection of a feature matrix that has the lowest number of features (18).

The PSO, Whale Optimization and Marine Predator algorithm give a good level of accuracy (above 90%), but the number of features are more than the Hybrid approach. In the next section, however, an examination of the machine learning models will be done to ascertain the performance of classifiers using these selected features.

TABLE VI. FEATURE SELECTION USING OPTIMIZATION

S.No	Optimization Algorithm	Accuracy	Number of selected features
1	Cuckoo's	92.5	40
2	LION	87.14	28
3	Particle Swarm Optimization	95.7	64
4	Marine Predators	92.1	22
5	Whale Optimization Algorithm	90.9	20
6	Cuckoo+LION (Hybrid)	98.9.0	18

IV. RESULT AND DISCUSSION

The accuracy of machine learning entirely depends on the quality of data it processes for learning patterns of data. This section explains the procedure followed for finally automating detecting four medical conditions of the eyes. For automation, four classifiers (KNN, SVM (Radial), DT, RF) were chosen based on the previous work done by other researchers and organisations.

The feature matrix of eighteen numerical features was selected using a hybrid feature selection algorithm (Cuckoo and LION), and it was subjected to all the four classifier

models. However, The rigorous experimentation showed that the accuracy of the SVM radial after full hyper parameter search and tuning give 95% accuracy as shown in Table VII. Correspondingly, the recall and precision values are also high. The better performance of the SVM radial algorithm can be attributed to the fact that the feature engineering process is paying off here. Secondly, to evaluate the consistency and validation of all classifier models, the ten-fold validation process was followed, and the standard deviation of each metric was noted. From Table VII and Table VIII, it can be observed that SVM radial have the lowest standard deviation for almost all the metrics, including recall and precision.

TABLE VII. PERFORMANCE ANALYSIS OF MACHINE LEARNING MODELS

Metric Algorithm	KNN	SVM (R)	DT	RF
Accuracy Cuckoo	0.84	0.88	0.84	0.80
Accuracy Lion	0.85	0.88	0.84	0.80
Accuracy Hybrid	0.89	0.95	0.83	0.83
F_ScoreCuckoo	0.87	0.89	0.83	0.82
F_ScoreLion	0.87	0.89	0.84	0.82
F_ScoreHybrid	0.87	0.91	0.83	0.82
PrecisionCuckoo	0.87	0.83	0.84	0.82
PrecisionLion	0.87	0.89	0.84	0.82
Precision Hybrid	0.87	0.91	0.83	0.83
Recall Cuckoo	0.80	0.84	0.83	0.82
Recall Lion	0.80	0.89	0.81	0.81
Recall Hybrid	0.80	0.91	0.84	0.85

TABLE VIII. STANDARD DEVIATION VALUES OF PERFORMANCE METRICS OF MACHINE MODELS

Metric Algorithm	KNN	SVM (R)	DT	RF
Accuracy Cuckoo	0.047589	0.041732	0.055656	0.055328
Accuracy Lion	0.044991	0.040143	0.053754	0.056553
Accuracy Hybrid	0.047383	0.021967	0.055517	0.054785
F_Score Cuckoo	0.036791	0.041807	0.054889	0.057068
F_Score Lion	0.037383	0.041967	0.03517	0.054785
F_Score Hybrid	0.037589	0.031732	0.035656	0.055328
Precision Cuckoo	0.034991	0.030143	0.033754	0.056553
Precision Lion	0.037383	0.031967	0.03517	0.034785
Precision Hybrid	0.036306	0.010105	0.03458	0.032732
Recall Cuckoo	0.037185	0.032468	0.034845	0.035144
Recall Lion	0.037383	0.031967	0.03517	0.034785
Recall Hybrid	0.037383	0.021967	0.03517	0.034785

*KNN=k-nearest Neighbours, SVM= Support Vector Machine, DT= Decision Tree, RF= Random Forest.

It can further be noted that the KNN algorithm is second best in terms of accuracy, and its performance metrics have higher levels of deviations compared to the SVM radial. Similar observations can be made for Decision tree and random forest algorithms. Both these algorithms have performed in a range of eighties per cent with higher levels of deviations in their results when evaluated for validations and reliability using the ten-fold method. It should be emphasised that the selection of these machine learning algorithms was made after conducting a bibliographic examination of the relevant literature. The methods that are most frequently employed to handle the challenges of classification and limited datasets have been incorporated into this book in their most basic forms. Because the dataset used in this study is an aggregate of various datasets, this research report includes a comparison of the dataset utilised in this study with the current dataset.

V. CONCLUSION AND FUTURE SCOPE

There have only been a few studies in which numerous medical eye problems have been investigated using a single method. The same can be said for identifying relevant picture features using a combinational technique. It was completed through a rigorous procedure that included a great deal of experimenting. The process of developing a generic pipeline of algorithms to facilitate feature selection and automation of the classification process has been followed to completion. An in-depth investigation of the optimization process was carried out in order to identify the most appropriate features and methods that could be used for the construction of the feature matrix, and further investigation resulted in the development of a numerically stable pipeline of the algorithms. It should also be mentioned that the KNN method is the second most accurate algorithm in terms of accuracy, and that its performance metrics have larger levels of deviations when compared to the SVM radial algorithm. Observations similar to these can be made about the decision tree and random forest algorithms. When examined for validation and reliability using the ten-fold approach, both of these algorithms performed within an eighty percent confidence interval, with larger degrees of variances in their results than when evaluated for accuracy.

Following a review of the literature on three keywords, the researchers chose the machine learning models for this work. The keywords were data fusion, eye illness classifiers, and image processing of the eyes. The procedure of picking the most accurate and stable classifier among the candidates was carried out with the assistance of a ten-fold algorithm, which was used to narrow down the field of candidates. This guaranteed that no time was wasted later on while assessing different machine learning models in the field. When hybrid algorithms (Cuckoo and LION) are used for feature engineering and dimension reduction, it has been discovered that there are extra benefits, and that this results in the generation of matrices with decreased features but complete coverage. This research was conducted under the guidance of an exploratory experimentation regime, and it has been discovered that the SVM radial algorithm is the most suited machine-learning model for the development of a multi-modality system that can detect eye abnormalities. In addition,

it was discovered that some degree of hyper-parameter adjustment was required in some cases. After conducting an extensive grid search based on hyper-parameter tuning and feature engineering, it was discovered that the optimization strategy resulted in a higher accuracy level (0.95) for SVM than the previous approach.

In this study project, we attempted to develop a multi-disease detection system that would be capable of detecting three different forms of eye diseases: diabetic retinopathy, glaucoma, and cataract, among others. It is recommended that other diseases be added to the scope in the future, and that the detection range be broadened as well. This way, the scalability and generality of the model can be further strengthened.

ACKNOWLEDGMENT

Authors are highly thankful to the RIC department of IKG Punjab Technical University, Kapurthala, Punjab, India for providing the resources and opportunity to conduct this research work.

DECLARATIONS FUNDING

The authors received no specific funding for this work.

CONFLICT OF INTEREST

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

AVAILABILITY OF DATA AND MATERIAL

The data that support the findings of this study is publically available on https://github.com/palavibhangu/retina_dataset

AVAILABILITY OF CODE

The code will be provided once the paper has been conditionally accepted.

REFERENCES

- [1] W. Y. Ng et al., "Blockchain applications in health care for COVID-19 and beyond: a systematic review," *Lancet Digit. Heal.*, 2021, doi: 10.1016/s2589-7500(21)00210-7.
- [2] A. Gupta and R. Katarya, "Social media based surveillance systems for healthcare using machine learning: A systematic review," *Journal of Biomedical Informatics*, vol. 108. 2020, doi: 10.1016/j.jbi.2020.103500.
- [3] T. F. Ursuleanu et al., "Deep learning application for analyzing of constituents and their correlations in the interpretations of medical images," *Diagnostics*, vol. 11, no. 8, 2021, doi: 10.3390/diagnostics11081373.
- [4] E. Omolara Abiodun et al., "A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities," *Neural Comput. Appl.*, vol. 33, doi: 10.1007/s00521-021-06406-8.
- [5] P. Dhar and M. Z. Abedin, "Bengali News Headline Categorization Using Optimized Machine Learning Pipeline," *Int. J. Inf. Eng. Electron. Bus.*, vol. 13, no. 1, pp. 15–24, Feb. 2021, doi: 10.5815/IJIEEB.2021.01.02.
- [6] S. Wang et al., "A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)," *Eur. Radiol.*, vol. 31, no. 8, 2021, doi: 10.1007/s00330-021-07715-1.
- [7] J. Carp, "On the plurality of (methodological) worlds: Estimating the analytic flexibility of fmri experiments," *Front. Neurosci.*, no. OCT, 2012, doi: 10.3389/fnins.2012.00149.

- [8] P. Kartikeyan and G. Shrivastava, "Review on Emerging Trends in Detection of Plant Diseases using Image Processing with Machine Learning," *Int. J. Comput. Appl.*, vol. 174, no. 11, 2021, doi: 10.5120/ijca2021920990.
- [9] H. Seo et al., "Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications," in *Medical Physics*, 2020, vol. 47, no. 5, doi: 10.1002/mp.13649.
- [10] A. Oniśko, P. Lucas, and M. J. Druzdzel, "Comparison of rule-based and Bayesian network approaches in medical diagnostic systems?," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2001, vol. 2101, doi: 10.1007/3-540-48229-6_40.
- [11] L. Oakden-Rayner, "Exploring Large-scale Public Medical Image Datasets," *Acad. Radiol.*, vol. 27, no. 1, 2020, doi: 10.1016/j.acra.2019.10.006.
- [12] Q. Abbas, "Glaucoma-Deep: Detection of Glaucoma Eye Disease on Retinal Fundus Images using Deep Learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, 2017, doi: 10.14569/ijacsa.2017.080606.
- [13] C. M. Sanders, S. L. Saltzstein, M. M. Schultzel, D. H. Nguyen, H. S. Stafford, and G. R. Sadler, "Understanding the limits of large datasets," *J. Cancer Educ.*, vol. 27, no. 4, 2012, doi: 10.1007/s13187-012-0383-7.
- [14] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, "A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2021, vol. 12614 LNCS, doi: 10.1007/978-3-030-71058-3_11.
- [15] G. Lim, V. Bellema, Y. Xie, X. Q. Lee, M. Y. T. Yip, and D. S. W. Ting, "Different fundus imaging modalities and technical factors in AI screening for diabetic retinopathy: a review," *Eye Vis.*, vol. 7, no. 1, 2020, doi: 10.1186/s40662-020-00182-7.
- [16] S. G. Francisco et al., "Dietary patterns, carbohydrates, and age-related eye diseases," *Nutrients*, vol. 12, no. 9, 2020, doi: 10.3390/nu12092862.
- [17] Y. Wang and S. Shan, "Accurate disease detection quantification of iris based retinal images using random implication image classifier technique," *Microprocess. Microsyst.*, vol. 80, 2021, doi: 10.1016/j.micpro.2020.103350.
- [18] G. R. Hemalakshmi, D. Santhi, V. R. S. Mani, A. Geetha, and N. B. Prakash, "Classification of retinal fundus image using MS-DRLBP features and CNN-RBF classifier," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 9, 2021, doi: 10.1007/s12652-020-02647-y.
- [19] O. Alfarraj, "A machine learning-assisted data aggregation and offloading system for cloud-IoT communication," *Peer-to-Peer Netw. Appl.*, 2020, doi: 10.1007/s12083-020-01014-0.
- [20] G. C. Ozmen et al., "An Interpretable Experimental Data Augmentation Method to Improve Knee Health Classification Using Joint Acoustic Emissions," *Ann. Biomed. Eng.*, vol. 49, no. 9, 2021, doi: 10.1007/s10439-021-02788-x.
- [21] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.
- [22] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, 2020, doi: 10.1016/j.inffus.2019.12.001.
- [23] J. Gao, P. Li, Z. Chen, and J. Zhang, "A survey on deep learning for multimodal data fusion," *Neural Computation*, vol. 32, no. 5, 2020, doi: 10.1162/neco_a_01273.
- [24] X. Wang, K. Wang, and S. Lian, "A survey on face data augmentation for the training of deep neural networks," *Neural Computing and Applications*, vol. 32, no. 19, 2020, doi: 10.1007/s00521-020-04748-3.
- [25] S. Uddin, A. Khan, E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," doi: 10.1186/s12911-019-1004-8.
- [26] J. Karthikeyan, S. H. P. Kumar, and K. Thirunavukkarasu, "Statistical techniques and tools for describing and analyzing data in Elt research," *Int. J. Civ. Eng. Technol.*, vol. 9, no. 11, 2018.
- [27] A. H. Halim, I. Ismail, and S. Das, "Performance assessment of the metaheuristic optimization algorithms: an exhaustive review," *Artif. Intell. Rev.*, vol. 54, no. 3, 2021, doi: 10.1007/s10462-020-09906-6.
- [28] A. Darwish, "Bio-inspired computing: Algorithms review, deep analysis, and the scope of applications," *Futur. Comput. Informatics J.*, vol. 3, no. 2, 2018, doi: 10.1016/j.fcij.2018.06.001.
- [29] S. R. Ahmad, A. A. Bakar, and M. R. Yaakub, "Ant colony optimization for text feature selection in sentiment analysis," *Intell. Data Anal.*, vol. 23, no. 1, 2019, doi: 10.3233/IDA-173740.
- [30] P. Kaur and R. Kumar Singh, "An Efficient Approach for Content-Based Image Retrieval Using Cuckoo Search Optimization," *Int. J. Model. Optim.*, vol. 9, no. 2, pp. 77–81, Apr. 2019, doi: 10.7763/ijmo.2019.v9.688.
- [31] L. Abualigah and A. Diabat, "A comprehensive survey of the Grasshopper optimization algorithm: results, variants, and applications," *Neural Computing and Applications*, vol. 32, no. 19, 2020, doi: 10.1007/s00521-020-04789-8.
- [32] S. S. Panicker and P. Gayathri, "Feature Selection Algorithms in Medical Data Classification: A Brief Survey and Experimentation," in *Lecture Notes in Electrical Engineering*, 2020, vol. 601, doi: 10.1007/978-981-15-1420-3_90.
- [33] K. K. Patro, A. Jaya Prakash, M. Jayamanmadha Rao, and P. Rajesh Kumar, "An Efficient Optimized Feature Selection with Machine Learning Approach for ECG Biometric Recognition," *IETE J. Res.*, 2020, doi: 10.1080/03772063.2020.1725663.
- [34] D. Devarajan, S. M. Ramesh, and B. Gomathy, "A metaheuristic segmentation framework for detection of retinal disorders from fundus images using a hybrid ant colony optimization," *Soft Comput.*, vol. 24, no. 17, 2020, doi: 10.1007/s00500-020-04753-7.
- [35] L. Parisi, N. RaviChandran, and M. L. Manaog, "Feature-driven machine learning to improve early diagnosis of parkinson's disease," *Expert Syst. Appl.*, vol. 110, 2018, doi: 10.1016/j.eswa.2018.06.003.
- [36] Sonali et al., "An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE," *Optics and Laser technology*, vol. 110, 2019, doi: 10.1016/j.optlastec.2018.06.061.
- [37] B. Sabir, F. Ullah, M. A. Babar, and R. Gaire, "Machine Learning for Detecting Data Exfiltration," *ACM Computing Surveys*, vol. 54, no. 3, 2021, doi: 10.1145/3442181.
- [38] V. T. S. M, A. Kumaravel, and K. B., "Gabor filter and machine learning based diabetic retinopathy analysis and detection," *Microprocess. Microsyst.*, 2020, doi: 10.1016/j.micpro.2020.103353.
- [39] M. Rai and P. Rivas, "A Review of Convolutional Neural Networks and Gabor Filters in Object Recognition," in *Proceedings - 2020 International Conference on Computational Science and Computational Intelligence, CSCI 2020*, 2020, doi: 10.1109/CSCI51800.2020.00289.
- [40] N. Sompairac et al., "Independent component analysis for unraveling the complexity of cancer omics datasets," *International Journal of Molecular Sciences*, vol. 20, no. 18, 2019, doi: 10.3390/ijms20184414.
- [41] M. Rostami, S. Forouzandeh, K. Berahmand, and M. Soltani, "Integration of multi-objective PSO based feature selection and node centrality for medical datasets," *Genomics*, vol. 112, no. 6, 2020, doi: 10.1016/j.ygeno.2020.07.027.
- [42] P. Kaur and R. K. Singh, "Content-based image retrieval using machine learning and soft computing techniques," *Int. J. Sci. Technol. Res.*, vol. 9, no. 1, 2020.

Educational Data Mining to Identify the Patterns of Use made by the University Professors of the Moodle Platform

Johan Calderon-Valenzuela, Keisi Payihuanca-Mamani, Norka Bedregal-Alpaca
Universidad Nacional de San Agustín de Arequipa
Arequipa, Perú

Abstract—Due to the events caused by the COVID-19 pandemic and social distancing measures, learning management systems have gained importance, preserving quality standards, they can be used to implement remote education or as support for face-to-face education. Consequently, it is important to know how teachers and students use them. In this work, clustering techniques are used to analyze the use, made by university professors, of the resources and activities of the Moodle platform. The CRISP-DM methodology was applied to implement a data mining process, based on the Simple K-Means algorithm; to identify associated groups of teachers it was necessary to categorize the data obtained from the platform. The Apriori algorithm was applied to identify associations in the use of resources and activities. Performance scales were established in the use of Moodle functionalities, the results show the use made by teachers was very low. Rules were generated to identify the associations between activities and resources. As a result the functionalities that need to be enhanced in the teacher training processes were identified. Having identified the patterns of use of the Moodle platform, it is concluded that it was necessary to use a Likert scale to transform the frequency of use of activities and resources and identify the rules of association that establish profiles of teachers and tools that should be promoted in future training actions.

Keywords—Clustering; educational data mining; moodle; usage patterns; k-means algorithm; a priori algorithm

I. INTRODUCTION

The pandemic caused by COVID-19 brought with it a mandatory social confinement that forced societies to change customs and models and to face the need to enhance the use of technological tools to give rise to teleworking and tele education.

The face-to-face educational system necessarily migrated towards virtual teaching-learning environments, which generated new ways of understanding the educational process. Integrating information and communication technologies (ICT) to the educational process implies changes in the forms of communication, in the contents and forms of evaluation, changes in the role of the teacher and students, ICT can be used by teachers as technical-pedagogical support and by students as a tool for autonomous learning [1].

Although some educational institutions were gradually integrating ICTs as a means of enhancing their educational processes, in the context of the pandemic, many of them had

to abruptly assume, without considering the context of their educational community, the use of virtual scenarios to carry out their educational activities.

Although some educational institutions were gradually integrating ICT as a means of enhancing their training processes, in the context of the pandemic, many of them had to assume abruptly and without considering the context of their educational community, the use of virtual scenarios to carry out their training activities. In particular, universities had to adapt to this transformation by assuming educational models that use technological tools of support and accompaniment to improve conventional teaching and learning processes. Within these tools are virtual educational platforms or LMS (Learning Management System) that enable new teaching-learning modalities. For [2] in the b-learning modality, which combines face-to-face with non-face-to-face teaching, the LMS favors learning. In the e-learning modality, the use of Internet-based technologies provides a wide range of solutions that promote the acquisition of knowledge and development of skills [3].

Training processes in virtual environments imply changes in the roles of their actors, the student stops being a passive consumer and becomes a producer of information and knowledge; for [4], the students must be autonomous and independent in their information search skills, must create new content in an innovative way and transform it into knowledge and must communicate effectively by decoding messages and transmitting information. For his part, the teacher assumes the role of guide and incorporates technology adequately and effectively, that is, incorporating technological tools with a pedagogical approach [5].

In this information context, in order for students to achieve meaningful learning, the teacher must use the LMS tools to manage resources and activities in a way that enhances the autonomous work of the student and favors the development of competencies and skills such as search and organization of information, teamwork and communication with their peers and with the teacher. That is why this research focuses on improving university teacher performance and consequently student learning.

For the analysis and evaluation of the use of LMS, data mining is a tool that allows determining patterns of behavior in the data obtained from the platform and identifies the factors associated with the success of online learning [6]. In

[7, 8] they found that most works focus on the usability of LMS, limiting themselves to the student's role and neglecting the teacher's work process.

In [9] a statistical analysis is made of how mathematics teachers in a face-to-face model supported by the Moodle platform develop cognitive and action competencies in elementary school students. In [10] educational data mining is applied to the Moodle LMS to identify behavior patterns in students to identify the resources and activities that are best suited to the students' needs; it concludes that there is a correlation between the level of activity and their academic performance.

In [11], the authors investigated the use of Virtual Learning Environments by professors in a higher education institution; they sequentially combined three methods: processing of VLE logs, surveys, and interviews. In [12] data is collected in some Spanish universities to study the uses that university professors make of the virtual campus and the methodology they propose to students. The model of courses carried out concludes that the universities studied make minimal use of the platform and focus on making materials available to the student and that the model is characterized by the presence of basic, complementary, and organizational teaching materials, together with a proposal of individual and group activities.

Works such as [13, 14, and 15] have used the CRISP-DM methodology to implement data mining processes aimed at describing students' academic behavior.

In [16], the authors use data mining tools and techniques for academic improvement of the student performance and to prevent drop out. Four classification methods, the J48, PART, Random Forest and Bayes Network Classifiers were used, the data mining tool used was WEKA.

The authors in [17], propose two different guidelines: Learning Analytics focused on descriptive processes, and Educational Data Mining for predictive processes, directing activities adjusted to this environment for obtaining satisfactory results.

In light of the above, the objective of this work was to analyze the use of resources and activities of the Moodle platform, made by teachers at the National University of San Agustín de Arequipa (Peru). The importance of the work carried out is that based on the results obtained, recommendations can be made to enhance the use of this platform and favor the use of interactive and collaborative activities within the framework of a socio-constructivist pedagogical model.

II. THEORETICAL FRAMEWORK

A. LMS Moodle

Modular Object-Oriented Dynamic Learning Environment (Moodle) is a learning management tool or LMS developed to create and manage online training environments. Known by its acronym, Moodle is one of the most widely used content management systems globally. It provides a powerful set of learner-centered tools and collaborative learning environments that empower both teaching and learning.

Moodle fosters active and participatory virtual environments by enabling teachers and students to interact on the platform through chats, forums, videoconferences, among others. Consequently, it facilitates to the teacher the possibility of extending the limits of the classroom to spaces and moments different from the face-to-face class, it gives students autonomy to consult multimedia content and to interact and participate in learning communities [18].

B. Educational Data Mining

Data mining is a set of techniques and technologies that, in an automatic or semi-automatic form, allows the exploration of large databases; the objective is to find repetitive patterns, trends, or rules that explain the behavior of the data in a specific context. Data mining involves the interaction of different techniques and procedures from computer science, statistics, mathematics, and information science. The extraction and analysis of data has become important in multiple areas, because through the application of various techniques, it allows to transform that data into information and knowledge of great utility [19].

In the educational field, there is talk of educational data mining (EDM); which focuses on the development of discovery methods that use data from LMS to understand and improve virtual teaching-learning environments. The EDM builds analytical models that uncover interesting patterns and trends in the use of an LMS [20].

C. CRISP-DM Methodology

CRISP-DM (Cross Industry Standard Process for Data Mining) is a standard and open analytical model of the data mining process [21]. It includes a description of the phases of a data mining project, the tasks required in each phase, and an explanation of the relationships between the tasks. CRISP-DM provides an overview of the life cycle of a data mining process.

CRISP-DM divides the data mining process into six main phases (Fig. 1).

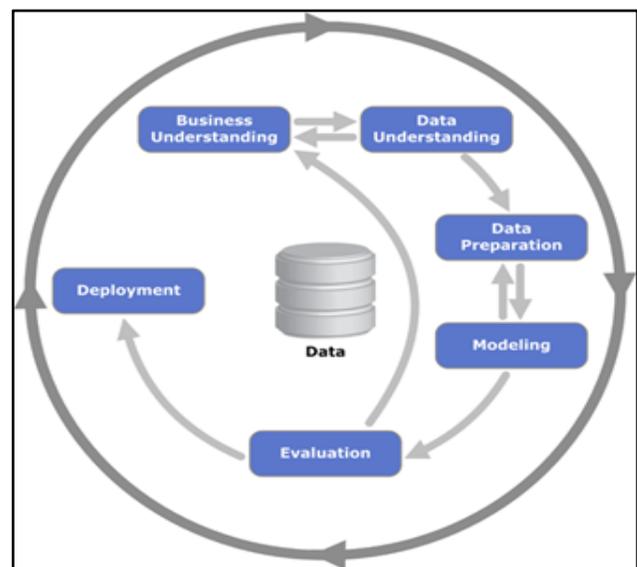


Fig. 1. CRISP-DM Phases. Source Shearer (2000).

The sequence of phases is dynamic; it is possible to move forward or backward through the phases. The result of each phase determines which phase, or which particular task of a phase, is to be done next. The arrows indicate the most important and frequent dependencies. The advantages of using CRISP-DM include replicability, independence of the application context, and its tool neutrality [22].

III. MATERIALS AND METHODS

A. Context

The research was conducted at the National University of San Agustín (UNSA) in Arequipa, Peru. Traditionally it works under a face-to-face model supported by the use of the Moodle platform. The University Directorate of Information Technologies (DUTIC-UNSA) is the unit in charge of the administration and management of the Moodle platform, in addition, it provides training and technical support services to teachers and students.

As a result of the COVID-19 pandemic, the UNSA migrated towards a mixed model in which face-to-face sessions are carried out through a videoconferencing system and arranged the mandatory use of the Moodle platform to dynamize and enhance the training process.

B. Methodology

Cross-sectional descriptive-exploratory research has been carried out.

To develop the data mining process, the steps of the CRISP-DM methodology have been followed.

C. Data

The study considered a universe composed of 4809 virtual classrooms implemented in the 2020-A academic semester. These classrooms correspond to the three academic areas of the UNSA: engineering, social and biomedical.

IV. DEVELOPMENT OF THE PROPOSAL

A. Phase 1, CRISP-DM Methodology: Understanding the Business

Tasks related to understanding project objectives and requirements were performed to turn them into technical objectives and a project plan.

Objectives: This research focuses on analyzing the use that teachers make of the activities and resources available on the Moodle platform. It is intended to:

- Investigate if there are differences in use between the three academic areas: Natural and Formal Sciences and Engineering, Social Sciences and Biomedical Sciences.
- Identify associated groups of teachers in the use of resources and activities using the Unsupervised Learning Algorithm Simple K-Means.
- Identify associations in the use of resources and activities using the A-priori association algorithm.

Resources:

- Technological: Computer, Google Collaboratory platform in Python language for data processing and model development.
- Technical: Data mining techniques.
- Human: The Researchers.
- Data source: Moodle platform usage logs during the 2020-A Academic Semester.

B. Phase 2, CRISP-DM methodology: Understanding the Data

A first contact with the data was made to familiarize themselves, identify their quality, define the first hypotheses and establish the most obvious relationships.

1) *Data collection*: The data was obtained from the records of the Moodle platform. For the data to have meaning it was necessary to integrate different tables. The "teacher" table with the correlative code of each teacher and their full name. The "course" table contains the correlative code of each virtual classroom, the short name and the full name of the corresponding subject. Both are related through the "course_teacher" table. The "resource_activities" table contains the number of times that the 13 activities and 7 resources available on the virtual platform have been used. This table relates to "course_teacher" through the course and teacher correlates. Fig. 2 shows the entity-relationship diagram of the integration performed.

2) *Verification of data quality*: Initially, the database considered the 4809 classrooms configured on the virtual platform. Data was cleaned considering that some classrooms had not been used or had not been assigned a teacher; 2 duplicate teacher records were deleted, 219 classrooms that had the Teaching field blank, and 335 classrooms that had less than or equal to a resource or activity used. As a result, we worked with data from 4255 classrooms.

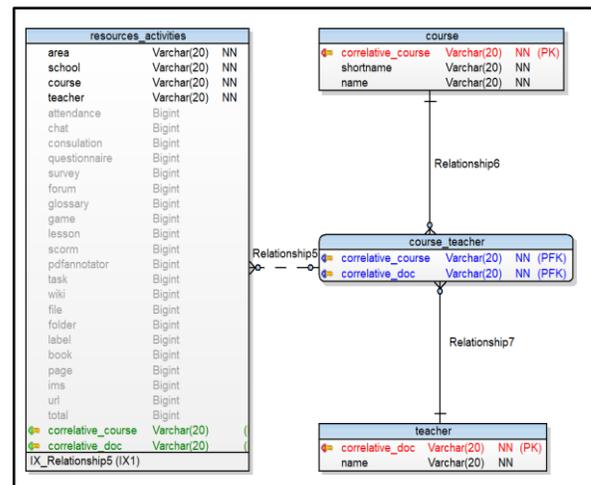


Fig. 2. Entity-relationship Diagram.

3) *Data exploration*: The use of classrooms was analyzed for each academic area, Fig. 3 shows the percentage of classrooms implemented and the number of teachers who were in charge in each of them. The differences found are explained by the number of professional schools that are integrated into each academic area.

At UNSA, the academic areas are divided into Schools and these into Professional Schools. Table I shows the number of virtual classrooms implemented in the different schools, as well as the number of professors administered them.

Tables II and III show the frequency of the use of Moodle resources and activities in each academic area and the amount of total usage.

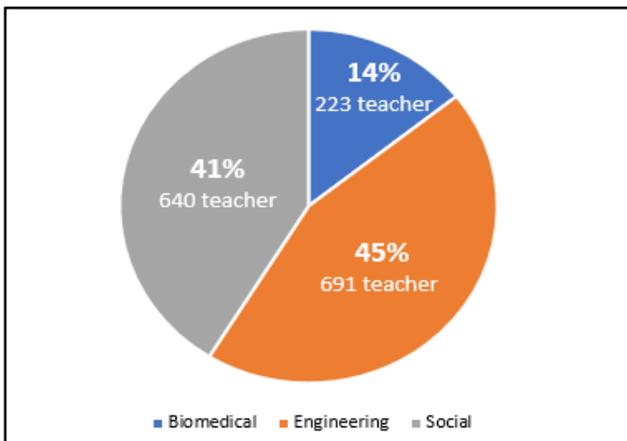


Fig. 3. Classrooms Percentage and Number of Professors in each Academic Area.

TABLE I. NUMBER OF CLASSROOMS BY SCHOOLS

School	No. of classrooms	No. of professors
School of Administration	269	115
School of Agronomy	93	48
School of Architecture	159	76
School of Biological Sciences	194	102
School of Accounting and Financial Sciences	172	80
School of Historical and Social Sciences	317	130
School of Natural and Formal Sciences	151	93
School of Law	189	77
School of Economics	130	49
School of Education	263	122
School of Nursing	76	47
School of Philosophy and Humanities	271	96
School of Geology, Geophysics and Mining	174	85
School of Civil Engineering	203	84
School of Process Engineering	403	210
School of Production and Services Engineering	750	313
School of Medicine	83	44
School of Psychology RRII Cs. Of Communication	358	146

TABLE II. FREQUENCY OF USE EACH RESOURCE BY AREA

Resources / Area	Biomedical	Engineering	Social	Total
File	13373	43299	46283	102955
Folder	675	1764	1512	3951
Label	2253	13042	9173	24468
Book	21	102	68	191
Page	149	224	238	611
Ims	0	2	0	2
Url	9476	33376	37343	80195
Total	16471	58433	57274	

TABLE III. FREQUENCY OF USE EACH ACTIVITY BY AREA

Resources / Area	Biomedical	Engineering	Social	Total
Attendance	1481	5421	5335	12237
Chat	275	1800	1614	3689
Consultation	10	81	106	197
Questionnaire	2771	6590	6697	16058
Survey	5	2	16	23
Forum	1469	4761	6962	13192
Glossary	71	87	138	296
Game	20	85	114	219
Lesson	19	287	176	482
Scorm	0	5	4	9
Pdf Annotator	152	169	117	438
Tasks	6808	21066	18617	46591
Wiki	14	29	39	82
Total	13081	40354	39896	

C. Phase 3, CRISP-DM Methodology: Data Preparation

In this phase, the necessary activities are carried out to build the data set that will serve for the modeling.

Each classroom was evaluated according to the frequency of use of each activity and resource. Finally, only the activities or resources used in at least 97% of the classrooms in each area were considered (Tables IV and V).

Consequently, in the academic area of Social Sciences ten tools were considered: Attendance, Chat, Questionnaire, Forum, Glossary, Task, File, Folder, Label and URL.

In the Engineering area, eleven tools were included: Assistance, Chat, Questionnaire, Forum, Glossary, Task, File, Folder, Label, Page and URL.

In the area of Biomedical Sciences, twelve tools were considered for the study: Attendance, Chat, Questionnaire, Forum, Glossary, Pdf Annotator, Task, File, Folder, Label, Page and URL.

In order to complement the information, Tables IV and V show the number of classrooms that have never used the activities or resources indicated by area of knowledge.

TABLE IV. NUMBER OF CLASSROOMS THAT HAVE NEVER USED EACH RESOURCE BY AREA

Resources / Area	Biomedical	Engineering	Social
File	10	47	34
Folder	337	1422	1545
Label	284	1198	1481
Book	437	1801	1932
Page	419	1774	1916
Ims	446	1838	1969
URL	46	327	265

TABLE V. NUMBER OF CLASSROOMS THAT HAVE NEVER USED EACH ACTIVITY BY AREA

Resources / Area	Biomedical	Engineering	Social
Attendance	173	1705	983
Chat	353	1781	1538
Consultation	436	1792	1916
Questionnaire	117	734	634
Survey	443	1838	1958
Forum	10	62	66
Glossary	416	1783	1881
Game	434	1803	1935
Lesson	433	1810	1925
Scorm	446	1835	1966
Pdf Annotator	429	1811	1918
Task	70	226	255
Wiki	437	1827	1948

Calculated the frequencies of use of the Moodle functionalities, it was found that the values were very dispersed so it was necessary to reduce them. Considering that scale transformations are efficient instruments for reducing a data set, it was decided to transform the values from numeric to nominal.

On the cleaned data, the maximum and minimum values of the frequency of use of each activity and resource were identified. With these values the range was calculated and divided into five parts of equal length, obtaining the values to construct a five-level Likert scale. The scale values and weight of each level assigned were:

- 0 = Very low
- 1 = Low
- 2 = Medium
- 3 = High
- 4 = Very high

D. Phase 4, CRISP-DM Methodology: Modeling

In this phase, the most appropriate modeling techniques are selected for the specific data mining project.

The Simple K-Means classification algorithm was applied, using the elbow method. It was looked the part of the graph

where the line changes abruptly which forms an "elbow"; that number of clusters will help when classifying the data. The appropriate was 30 iterations to determine the optimal number of clusters, obtaining $k = 5$ (Fig. 4).

The objective was to group similar observations and discovers patterns in the use of Moodle resources and activities in the transformed data based on the Likert scale.

It was worked with 4255 classrooms; each classroom has 30 fields including: area, faculty, school, course, teacher, activities and resources. Each classroom has the amount of resources and activities used numerically, therefore the Likert scale was applied to have a standard score from 0 to 4 with respect to usability; then, we eliminate columns that are not going to be trained, the data set used consists of 20 fields between activities and resources, which will be grouped using the K means algorithm, resulting in 5 groupings, cluster 1, 2 and 4 have less usability, cluster 5 has a medium usability followed by cluster 3 (Fig. 5) and Table VII.

To identify associations in the use of resources and activities, the A-priori association algorithm was applied in each academic area.

First, we converted the data into 0's (activities and resources with score 0) and 1's (activities and resources with score 1,2,3 and 4) to work the model.

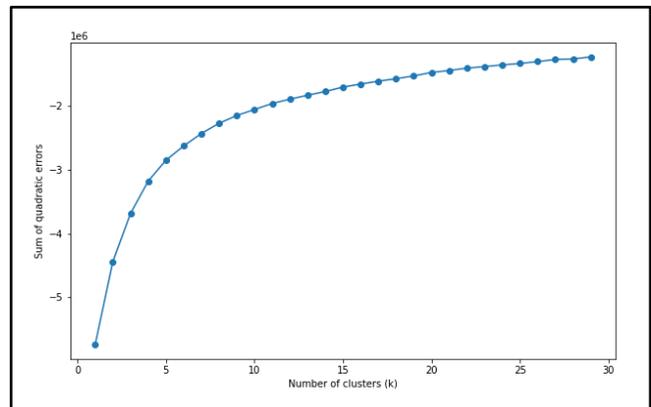


Fig. 4. Elbow Method to Determine the Optimal value of k.

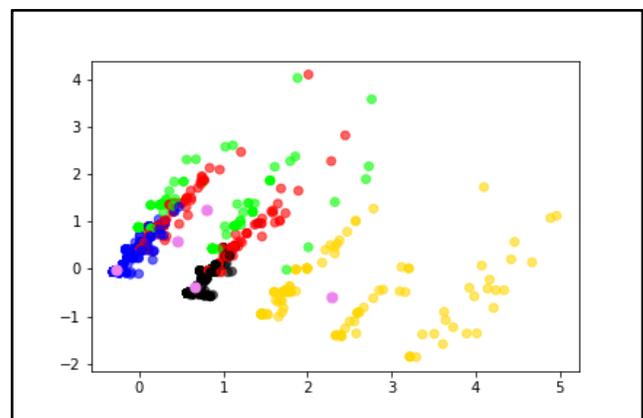


Fig. 5. Clusters Determined by Simple K-Means.

In the social area, associations were obtained with the tools file, URL, and task. In the biomedical area, associations were obtained with the questionnaire, homework, URL, and attendance. However, in the engineering area no association was found; therefore. Due to these differences, it was decided to perform the process again with the total data set.

Considering the definition of support given by [23], the association rules were constructed with a minimum support of 0.8; increasing the number of associations, where the following is true: antecedent = is used => consequent = is used, the probability of each association is observed in the "Confidence" column, the Lift column shows the increase in the probability of use when considering both tools rather than only the antecedent tool. (Table VI shows the results).

TABLE VI. ASSOCIATIONS FOUND WITH THE APRIORI ALGORITHM

Antecedent	Consequent	Support	Confidence	Lift
URL	file	0.840	0.988	1.009
forum. URL	file	0.810	0.987	1.009
task	file	0.859	0.987	1.009
task, forum	file	0.830	0.987	1.008
task	file, forum	0.830	0.954	1.008
URL	file, forum	0.810	0.953	1.008
File	task	0.859	0.878	1.009
file, forum	task	0.830	0.877	1.008
file	URL	0.840	0.858	1.009
file, forum	URL	0.810	0.856	1.008
file	file, forum	0.830	0.848	1.008
file	forum, URL	0.810	0.828	1.009

V. RESULT AND DISCUSSION

When analyzing Tables II and III, it can be affirmed that the resource "File" has been used primarily, followed by the resource "URL"; it follows then, that the virtual classroom is being used primarily as a repository of content and links to web content. Likewise, the activity "Homework" has been mostly used, followed by the activity "Questionnaire"; which means that the virtual classroom is being used to collect assignments and make assessments through questionnaires.

Table VIII (at the end of the paper) shows the results obtained by applying the Simple K-Means algorithm to the processed data. The nominal values of the Likert scale were considered for each of the tools in the clusters, since the distribution of the frequency of use is asymmetric, the value of the model was taken as a representative measure. Table VIII also shows that the teachers' use of Moodle activities and resources was classified with very low performance. Although this situation differs slightly in the areas of Social Sciences and Biomedical Sciences, it was found that in Engineering there was not enough support to find association rules.

There are no reasons to assume that teachers do not know how to use the different functionalities of the virtual platform; however, these results suggest the need to reinforce teacher training in the use of the activities and resources provided by Moodle. The training process should focus on the use of the virtual platform as a support to the synchronous sessions conducted by the teacher so that learning outcomes are enhanced through the implementation of active teaching-learning methodologies that can be easily implemented with the functionalities provided by Moodle [24,25,26].

Table VII shows the association rules obtained as a result of applying the Apriori algorithm to the cleaned data corresponding to the three academic areas. The association rules relate to the use of Moodle activities and resources.

TABLE VII. ASSOCIATION RULES OBTAINED FROM THE APRIORI ALGORITHM

#	Rules of association	Interpretation in Natural language
1	URL = is used=> file= is used	If the URL tool is used in the classroom, then the file tool will be used.
2	forum, URL = is used => file = is used	If both the forum and URL tools are used in the classroom, then the file tool will be used.
3	task = is used => file = is used	If in the classroom the task tool is used, then the file tool will be used.
4	task, forum = is used => file = is used	If both task and forum, are used in the classroom, then the file tool will be used
5	task = is used=> file, forum = is used	If in the classroom the task tool is used, then the file and forum tools will be used.
6	URL = is used => file, forum = is used	If the URL tool is used in the classroom, then the file and forum tools will be used.
7	file = is used => task = is used	If the file tool is used in the classroom, then the task tool will be used.
8	file, forum = is used => task = is used	If in the classroom the tools file and forum are used, then the task tool will be used.
9	file = is used => URL = is used	If the file tool is used in the classroom, then the URL tool is used
10	file, forum = is used => URL = is used	If both the file and forum tools are used in the classroom, then the URL tool will be used.
11	file = is used => task, forum = is used	If the file tool is used in the classroom, then the task and forum tools will be used.
12	file = is used => forum, URL = is used	If in the classroom the file tool is used, then the forum and URL tools will be used.

Association rules have been generated and they have allowed the identification of associations between activities and resources, rules that give evidence of the activities and resources that need to be enhanced in the teacher training processes. For example, from rules 5, 6, and 11 it can be inferred that if the teacher uses the task activity and the resources file and forum, then it is very likely that he/she will implement discussion forums among the activities to be implemented in the virtual classroom; therefore, he/she will not only be imparting knowledge but also supporting the development of communication skills and critical thinking.

TABLE VIII. RESULTS OF THE SIMPLE K-MEANS CLUSTERING ALGORITHM

Variable	Group 1 (3291 classrooms)		Group 2 (72 classrooms)		Group 3 (89 classrooms)		Group 4 (621 classrooms)		Group 5 (182 classrooms)	
Attendance	3231	Very low	65	Very low	66	Medium	595	Very low	160	Very low
Chat	3265	Very low	64	Very low	88	Very low	594	Very low	178	Very low
Consultation	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
Questionnaire	3254	Very low	55	Very low	79	Very low	583	Very low	122	Very low
Survey	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
Forum	3232	Very low	60	Very low	88	Very low	603	Very low	163	Very low
Glossary	3280	Very low	67	Very low	87	Very low	610	Very low	176	Very low
Game	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
Lesson	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
Scorm	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
Pdf Annotator	3285	Very low	69	Very low	89	Very low	621	Very low	182	Very low
Task	3291	Very low	26	Very low	45	Very low	549	Very low	118	Very low
Wiki	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
File	3140	Very low	38	Very low	66	Very low	537	Very low	109	Low
Folder	3274	Very low	62	Very low	82	Very low	604	Very low	177	Very low
Label	3230	Very low	62	Very low	82	Very low	582	Very low	158	Very low
Book	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
Page	3291	Very low	71	Very low	89	Very low	616	Very low	182	Very low
Ims	3291	Very low	72	Very low	89	Very low	621	Very low	182	Very low
url	3158	Very low	31	Very low	59	Very low	553	Very low	103	Low

VI. CONCLUSION

With the development of this work, it was possible to identify the behavior patterns of university professors in the use of the activities and resources offered by the Moodle platform.

The use of a Likert scale to transform the frequency of use of activities and resources allowed to reduce the spectrum of values and to be able to find associations when applying the K-Means algorithm; using the elbow method with 30 it was determined that the optimal was to work with 5 clusters.

The activity of the teachers was characterized and it was found that the activities: chat, wiki, lesson, workshop, questionnaire, games and survey are not used despite their great potential as didactic material that can enhance the results of the teaching and learning processes.

The results obtained in this work will serve to implement teacher training processes focused on the proper use of Moodle activities and resources that allow the development of virtual or blended courses based on constructionist and social constructivist approaches.

Applying data mining techniques to the large amount of information generated by the Moodle platform can contribute to the creation of dynamic profiles in the development of a virtual course. In addition to improvements in the teacher's use of resources and activities, students' behavior patterns could be considered in order to adapt courses to their level of learning.

REFERENCES

- [1] I. Martí Castro, "Aprendizaje-Virtual". En Diccionario Enciclopédico de Educación. Grupo Editorial Ceac S. A. (LEXUS), 2003.
- [2] A. Bartolomé, "Blended Learning. Conceptos básicos. Pixel-Bit". En Medios y Educación, 23. Pp. 7-20. [en línea], 2004.
- [3] D. Alemany Martínez, "Blended learning: modelo virtual – presencial de aprendizaje y aplicación en entornos educativos". España, Alicante: Universidad de Alicante, 2010.
- [4] J. Cabero, "La formación en la sociedad del conocimiento. INDIVISA - Boletín de Estudios E Investigación - Monografía X: Las TICs en los contextos de formación universitaria". 13-48, 2008.
- [5] F. Pedró, "Tecnología y escuela: lo que funciona y por qué: Documento Básico". Madrid: Fundación Santillana. Obtenido de http://www.fundacionsantillana.com/upload/ficheros/noticias/201111/documento_bsico.pdf, 2011.
- [6] R. Herrera, Myriam; Ruiz, Susana; Romagnano, María; Ganga, Leonel; Lund, María Inés y Torres, Estela. "Aplicando métodos y técnicas de la ciencia de los datos a datos universitarios". XXI Workshop de Investigadores en Ciencias de la Computación, Facultad de Ciencias Exactas, Físicas y Naturales, Universidad Nacional de San Juan, Argentina, Abril 2019.
- [7] A. Van Leeuwen; J. Janssen; G. Erkens; M. Brekelmans, "Supporting teachers in guiding collaborating students: Effects of learning analytics in CSCL". Computers & Education, vol. 79, 28-39, 2014.
- [8] J. M. Doderó, F. J. García-Peñalvo, "Development of E-Learning Solutions: Different Approaches, a Common Mission". Revista Iberoamericana de Tecnologías del Aprendizaje, IEEE-RITA vol. 9(2), 72-80 et al. 2014.
- [9] G. Hernández, "Análisis del uso y manejo de la plataforma Moodle en docentes de matemáticas, para el desarrollo de competencias integrales en estudiantes de primaria". Revista Q, 10 (19). <http://dx.doi.org/10.18566/revistaq.v10n19.a01>, 2015.
- [10] B. Hidalgo Cajo, "Minería de datos en los Sistemas de gestión de Aprendizaje en la Educación Universitaria". Campus Virtuales, ISSN-e 2255-1514, Vol. 7, N° 2, 2018.

- [11] G. Samaniego, L. Marqués, y M. Gisbert, "El profesorado universitario y el uso de Entornos virtuales de aprendizaje". Campus Virtuales, Vol. IV, Num. 2, pp. 50-58, 2015.
- [12] Salinas, "Modelos didácticos en los campus virtuales universitarios: Perfiles metodológicos de los profesores en procesos de enseñanza-aprendizaje en entornos virtuales". Reporte de investigación, Enero, Conferencia Virtual Educa, 2008.
- [13] A. I. Oviedo Carrascal, J. Jiménez Giraldo, "Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas" SABER-PRO. Revista Politécnica, 15(29), 128-140. <https://doi.org/10.33571/rpolitec.v15n29a10>, 2019.
- [14] N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, P. Yanque-Churo, "Classification models for determining types of academic risk and predicting dropout in university students". *International Journal of Advanced Computer Science and Applications (IJACSA)*, Volume 11 Issue 1, 2020.
- [15] D. Buenaño-Fernández, "Uso de la metodología CRISP-DM para guiar el proceso de minería de datos en LMS". In book: Tecnología, innovación e investigación en los procesos de enseñanza aprendizaje Rosabel Roig-Vila (Ed.) Chapter: Uso de la metodología CRISP-DM para guiar el proceso de minería de datos en LMS. Editors: Ediciones Octaedro, 2016.
- [16] S. Hussain, N. Abdulaziz, F. Ba-Alwi & N. Ribata. "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol. 9, No. 2, February 2018, pp. 447~459, ISSN: 2502-4752, DOI: 10.11591/ijeecs.v9.i2.pp447-459.
- [17] J. Salazar-Cardona & J. Triviño-Arbeláez, "Aplicación de learning analytics y educational data mining en una institución de educación superior en Colombia", *Revista Ingenierías Universidad de Medellín*, 19(36) • Enero-Junio de 2020 • pp. 71-89 • ISSN (en línea): 2248-4094. DOI 10.22395/rium.v19n36a4.
- [18] I. Peña, C. Córcoles, & C. Casado, "El profesor 2.0: docencia e investigación desde la red". *Revista sobre la sociedad del conocimiento*, 3, 1-9. Obtenido de <http://ullviu.blog.cat/gallery/12161/12161-68417.pdf>, 2006.
- [19] B. Martín Galán, & D. Rodríguez Mateos, "La evaluación de la formación universitaria semipresencial y en línea en el contexto del EEES mediante el uso de los informes de actividad de la plataforma Moodle". *RIED. Revista Iberoamericana De Educación a Distancia*, 15(1), 159-178. <https://doi.org/10.5944/ried.1.15.782>, 2012.
- [20] R. Cristóbal, S. Ventura, E. García, "Data mining in course management systems: Moodle case study and tutorial". *Computers and Education*, 51(1), 368-384. doi: <https://doi.org/10.1016/j.compedu.2007.05.016>, 2008.
- [21] C. Shearer, "el modelo CRISP-DM: el nuevo plan para la minería de datos, almacenamiento de los datos". J 5:13-22, 2000.
- [22] J.M. Moine, A.S. Haedo, & S. Gordillo, "Estudio comparativo de metodologías para minería de datos". *CACIC 2011 - XVII CONGRESO ARGENTINO DE CIENCIAS DE LA COMPUTACIÓN*, 2011.
- [23] K. C. Mondal, B. D. Nandy, & A. Baidya, "A Factual Analysis of Improved Python Implementation of Apriori Algorithm". *Methodologies and Application Issues of Contemporary Computing Framework*, 139-151, 2018.
- [24] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaén, V. Cornejo-Aparicio. "Video and Cooperative Work as Didactic Strategies to Enrich Learning and Development of Generic Competences in numerical Methods". *XIII Latin American Conference on Learning Technologies (LACLO)*. 2018. DOI: 10.1109/laclo.2018.00038.
- [25] N. Bedregal-Alpaca, D. Tupacyupanqui-Jaen, M. Rodríguez-Quiroz, L. Delgado-Barra, K. Guevara-Puente and O. Sharhorododka, "Problem-Based Learning with ICT Support: An experience in teaching-learning the concept of derivative," 2019 38th International Conference of the Chilean Computer Science Society (SCCC), Concepcion, Chile, 2019, pp. 1-7. DOI: 10.1109/SCCC49216.2019.8966396.
- [26] N. Bedregal-Alpaca, V. Cornejo-Aparicio, A. Padron-Alvarez, E. Castañeda-Huaman, "Design of cooperative activities in teaching-learning university subjects: Elaboration of a proposal", *International Journal of Advanced Computer Science and Applications* Volume 11, Issue 42020 Article number 0445. DOI 10.14569/IJACSA.2020.0110445.

Assessing the Quality of Educational Websites in Sudan using Quality Model Criteria through an Electronic Tool

Asim Seedahmed Ali Osman
College of Computer Science and Engineering
University of Hafr Al Batin, Hafar Al Batin, Saudi Arabia

Abstract—The application of internet has grown in recent years. Due to this, there is an increase in the number of websites which creates diversity in the services. This leads the researchers to do more research in the quality of websites in order to set standards and models to maintain its quality. The main objective of these standards is to support the trust and speed which is the cornerstone which are the basis for using websites. Various statistics reports show that the sites of institutions and companies which applied the quality standards have achieved high rates in terms of the user satisfaction and the number of visitors. This study has incorporates the concept of website and electronic gates, its objectives, advantages, types and in addition to its quality and standards of e-websites. It also touched on previous studies conducted on website in the world, the Arab world and African peninsula and also in the Sudan in support of the development of Sudanese websites. The proposed models consist of important metrics to evaluate the application, quality of content, aesthetic aspects, multimedia, reputation and security etc. This paper also proposed an application for evaluating the quality of websites based. This model is applied on Sudanese websites such as governmental, educational, and commercial etc. The authors used the object oriented programming approach to build the proposed model using the PHP language with the combination of CSS and Java script.

Keywords—Website quality; e-websites; education; information technology; PHP language

I. INTRODUCTION

Information Technology plays an important and tangible role in our scientific life. Many institutions have adopted the use of it and are considered as a mediator in the learning and communication process in universities. This improves and develops the product of education and its implementation for various programs or sites. This also helps the students to join these educational institutions. In this paper, a new methodology is proposed to evaluate the various educational website in different universities present inside the Sudan.

The educational website of various universities contains different entities such as the main pages of various site, the glimpse about the respective university, its vision, mission, various colleges and its departments etc [1]. It also consists of main page of the course, information about the course, information about the faculty members who are working in the particular college, the announcements and video meetings, topics related to homework, study materials and activities,

course content and electronic reference list, information about virtual classes, events done in the conference rooms and the explanations about synchronous and asynchronously communication etc.

The universities faced various problems during the instance of providing the important components of website, especially at the process of evaluating the educational websites of Sudanese universities at the international level [3]. The evaluation process may be carried out low often due to the lack of control and failure to considering some requirements of quality. This is due to the non-application of international quality standards such as (ISO),(OSI) [4] which are the set of internationally recognized technical specifications used for the purpose of evaluating the websites. It can be done in order to operate the production and commodity processes, models, performance and management, which includes the esthetically aspects etc. Content Quality Standard is done in the search engine, information guide, maps and automatic updating for site where, the site also requires a standard Multimedia which is represented by video, audio and motion systems which make the site more attractive, smoother interaction, smoother in clarify the content etc.

The main aim of this research it is study the analysis part of the various Sudanese educational websites [4] using different quality standards for the purpose of assessing the quality of educational sites and along with its services. This paper also indents to develop a comprehensive framework model which contains all the main quality elements and accompanying indicators for accessing a quality of website discussed in previous researches. Based on the previous researches, we proposed this paper to achieve the following goals:

- To define of quality standards for educational website.
- To define the standards and quality models for educational websites.
- To apply various standards of quality to Sudanese educational websites.
- To disseminate and support the Sudanese e-government and to perform its services with high efficiency and interactivity.

- To propose a model for evaluating the quality of educational websites based on appropriate standards and measurements for determining its quality.
- To improving the work environment and to change the quality of performance of Sudanese educational websites.
- To improving the confidence of customers using these educational websites.
- To make decisions based on real information from educational websites.
- To create websites in an integrated way to develop the product of education.

In order to achieve these goals, it is necessary to design a form for evaluating the educational websites [7] by knowing the performance quality of these educational websites of Sudanese universities. The form allows the entry of an educational website URL for the university and the evaluation is carried out by the user and the system administrator by clicking the evaluation button. The site is evaluated based on the calculations and equations of the quality criteria that have been mentioned. It shows the evaluation result and the outcome of each websites evaluation in addition to a graph showing values of criteria. There are also details about the reports of evaluation when selecting evaluation details of websites.

Remainder of the paper is summarized as follows. Section 2 depicts the review of various methods for accessing the quality of normal websites and also the educational websites. Section 3 provides an attribute of the proposed methodology with its functional architecture and its working principle. Section 4 explains the various criteria for evaluating the quality of website along with its working principle and various formulas. Section 5 proposes the evaluation results for various university websites present in the Sudan. Section 6 shows the conclusion of this paper.

II. LITERATURE REVIEW

Recently, the internet has created a new environment enabling any organization to conduct its entire set of processes and practices of business especially through online [1]. E-learning and its associated activities is any process performed via an Internet-based, computer-mediated network [2]. There are many categories of e-learning which tend to be used interchangeably leading to policy incoherence. Majority of these learning practices were done by the application of Websites.

Web based artifacts have been developed over the past few years. The increasing acceptability of websites increases the challenges, for instance, in knowing or assessing where we are standing regarding the quality of product, and how it can be improved [3]. However, there is still no methodology which is widely recognized as quantitative process for the evaluation of quality of websites. One of the main goals in the evaluation of website is qualitative and comparison based process. This is done in order to understand the entire process which a set of quality characteristics and attributes etc. Various methods

have been proposed by earlier researchers [4-10, 20] in this process.

Authors in [11] proposed a quantitative evaluation approach to assess the websites quality called Website Quality Evaluation Method (QEM). This approach might be useful to evaluate and compare characteristics quality and attributes of a Web product lifecycle in different phases. Particularly, the authors evaluated their proposed methodology in six academic websites sites. At the end of the evaluation process, a ranking is obtained for each selected site. Specifically, the evaluation process generates elemental, partial, and global indicators or quality preferences that can be easily analyzed, backward and forward traced, justified, and efficiently employed in decision-making activities.

Educational websites are analyzed in different perspectives. Authors in [12] developed a theoretical based framework for evaluating the quality of website from the perspective of user satisfaction. Other researchers concentrated on some specific features of websites such as authors in [13] developed a framework to measure the importance of usability of websites, while authors in [14] investigated and evaluated the design of university websites. Other researchers, while assessing the websites of universities considered various other features. Authors in [15] designed criteria to evaluate the resources of websites for maximum utilization of the scholarly context and research within the area of the art and history. Authors in [16, 19] tried to find solutions to problems among the user and involved evaluating various university websites inside the South African based on certain factors.

Author in [2] emphasized the learning practices by the application of various websites. Authors in [1] stress the importance of E-learning and its associated activities through the intern and on-line based applications. Website Quality Evaluation was done by [11] where, the characteristics quality and attributes of a Web product lifecycle in different phases were evaluated and compared. It can also be used in the decision making activities. User satisfaction is considered by authors in [12]. They proposed a theoretical based framework for evaluating the quality of website based on the user satisfaction. Usability of websites was considered and a framework was developed for it by authors [13]. Their framework also empathizes to define the standards and quality models for various educational websites. The main drawback behind their framework is applied only for the educational websites. Authors in [16, 19] found various solutions to problems among the user of various university websites inside the South African based on certain factors. Main drawbacks of their works is it is applies only inside the South African University websites.

III. PROPOSED METHODOLOGY

Authors developed a standard which depends on various ISO based Quality model for assessing the quality of educational websites. It is divided into six levels based on their different purposes and topics. It is further directed to evaluate the sites particularly belongs to the Sudanese educational institutions. Various criteria proposed in [18] were applied in the proposed model. The proposed quality model

shown in Fig. 1 includes various metrics which represents the frameworks for determining the quality of the educational website in terms of its content, structure and presentation. These metrics are as follows:

- Aesthetics
- Ease of Use
- Rich Contents
- Multimedia
- Reputation
- Security

A. Aesthetics

Aesthetic aspects of an element maintain a fundamental importance in evaluating the quality of websites since, it leaves a positive impact for the visitor of the educational website. It works on by monitoring the interaction between the user and the educational website. It also monitors some aspects of the Human-Computer-Interaction (HCI) [17] branch and also by the incentives to interact with the site such as availability which includes:

- Image: 40%.
- Colors taken 20%.
- Tables and text, backgrounds, etc.
- Resolution and standard Table takes 40%.

The aesthetics scale is calculated based on the following equation:

$$\text{Aesthetics} = \text{images} * 0.4 + \text{tands} * 0.2 + \text{color} * 0.2 + \text{underline} * 0.2 \quad (1)$$

B. Ease of Use

One of the most important problems in the factor for ease of use is the measurement of Usability is for websites or for the products of software etc. Various criteria or models have been proposed earlier for measuring and assessing the usability within societies interacting between human being and the software engineering based societies.

This metric contains of various sub-metrics like:

- The Navigation offer consists of 30%.
- Annotation alerts and routers consist of 30%.
- CSS consists of 40%.

The scale for ease of use is calculated based on the following formula:

$$\text{Ease of Use} = \text{cssl} * 0.4 + \text{nav} * 0.3 + \text{ann} * 0.3 \quad (2)$$

C. Rich Contents

The quality of the website contents can be described by the availability of various elements such as:

- Search Engine, which takes 20%.
- Graphs, which takes 30%.
- The Information Guide, which takes 30%.
- Avoiding Auto-refresh, which takes 20%.

Other elements that make the website more responsive to the requests of visitors and users fall under the content quality scale. It can be calculated based on the following formula.

$$\text{Rich Contents} = \text{bulletin} * 0.3 + \text{guide} * 0.3 + \text{searchenginge} * 0.2 + \text{autorefresh} * 0.2 \quad (3)$$

D. Multimedia

Multimedia represents the video and audio elements, movement systems etc. When these elements were used inside the site, it will make the website more attractive, interactive and smoother in terms of working. The multimedia elements such as sound and image can be appropriate as content of the site. This can be as follows:

- One Media in one Page takes 30%.
- Using thumb mails takes 30%.
- Attributes of multimedia and its components take 10%.
- Plug-in support takes 30%.

The vision differs from the aesthetic aspects and ease of use of application. The multimedia scale is calculated based on the following formula:

$$\text{Multimedia} = \text{plugin} * 0.3 + \text{thumbnail} * 0.3 + \text{attribute} * 0.1 + \text{minone} * 0.3 \quad (4)$$

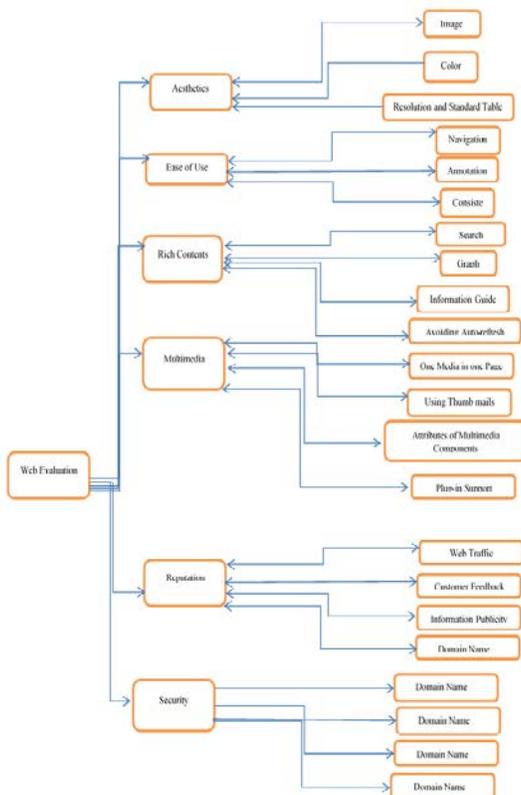


Fig. 1. Various Criteria of the Proposed Model.

E. Reputation

The reputation of a website defines the extension of reliability of it and the extent of interaction which is accepted by the environment for which it is directed or by the local, regional or international bodies. These are accredited to measure the quality of the websites which covers the following:

- Web Traffic efficiency takes 30%.
- Customer Feedback takes 30%.
- Information and Publicity takes 20%.
- Domain Name takes 20%.

The reputation scale is calculated based on the following formula:

$$\text{Reputation} = 0.3 * \text{feedback} + 0.2 * \text{domain} + 0.3 * \text{traffic} + 0.2 * \text{publicity} \quad (5)$$

F. Security

Information security plays an important role in the protection of assets of an institution. We often hear about security incidents of information security, such as site distortion, server piracy, and data leakage etc. Hence, there is an urgent need to devote more resources to protecting information assets. It is as follows:

- The login mechanism takes 40%.
- The firewall takes 30%.
- User display levels and usage of session variables when interacting with the site takes 30%.

The security scale is calculated based on the following formula

$$\text{Security} = 0.3 * \text{firewall} + 0.3 * \text{sesion} + 0.4 * \text{login} \quad (6)$$

IV. IMPLEMENTATION OF THE EVALUATION REPORT SCREEN

Following are the various outputs of evaluation reports done in different types of Sudanese educational websites. Implementation of the evaluation details screen is shown in Fig. 2 which is one of the implementation result of the proposed system.



Fig. 2. Implementation of the Evaluation Details Screen.

The above figure depicts the implemented image of the evaluation details screen with all the details of the evaluation of the website, details of all standards such as aesthetics and multimedia, quality of content, reputation, ease of use and security etc.



Fig. 3. Implementation Screen of the Content Quality Details.

Fig. 3 shows the implementation of the content quality and the details screen. It has various criteria for measuring the quality of the content in addition to the value of the criterion in the evaluation of result of the actual website.

V. EVALUATION RESULTS OF VARIOUS EDUCATIONAL WEBSITES

The proposed system is tested for the various Sudanese educational websites. The following diagrams represent the evaluation results of the various educational websites. Fig. 4 evaluation results of (www.neelain.sd), Fig. 5 evaluation results of (www.sustech.edu), Fig. 6 evaluation results of (www.uofk.edu), Fig. 7 evaluation results of (www.uofg.edu.sd), Fig. 8 evaluation results of (www.aau.edu.sd), Fig. 9 evaluation results of (www.siu-sd.com), Fig. 10 evaluation results of (www.bahri.edu.sd), Fig. 11 evaluation results of (www.iau.edu.sa), Fig. 12 evaluation results of (www.oiu.edu.sa), Fig. 13 evaluation results of (www.mu.edu.sd), Fig. 14 evaluation results of (www.usd.edu.sd) and Fig. 15 evaluation results of (www.ribat.edu.sd).

0.37	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.5	السمعة
1	سهولة الاستخدام
0.3	الأمنية
0.62	النتيجة
جدد	نتيجة التقييم

Fig. 4. www.neelain.edu.sd.

0.47	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.5	السمعة
1	سهولة الاستخدام
0.3	الأمنية
0.63	النتيجة
جدد	نتيجة التقييم

Fig. 5. www.sustech.edu.

0.3	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.2	السمعة
0.6	سهولة الاستخدام
0.3	الأمنية
0.43	النتيجة
ضعيف	نتيجة التقييم

Fig. 6. www.uofk.edu.

0.47	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.5	السمعة
1	سهولة الاستخدام
0.7	الأمنية
0.71	النتيجة
جيد جدا	نتيجة التقييم

Fig. 12. www.oiu.edu.sa.

0.47	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.5	السمعة
1	سهولة الاستخدام
0.3	الأمنية
0.63	النتيجة
جيد	نتيجة التقييم

Fig. 7. www.uofg.edu.sd.

0.47	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.2	السمعة
1	سهولة الاستخدام
0.7	الأمنية
0.65	النتيجة
جيد	نتيجة التقييم

Fig. 13. www.mu.edu.sd.

0.27	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.2	السمعة
1	سهولة الاستخدام
0.3	الأمنية
0.55	النتيجة
جيد	نتيجة التقييم

Fig. 8. www.aau.edu.sd.

0.37	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.5	السمعة
1	سهولة الاستخدام
0.3	الأمنية
0.62	النتيجة
جيد	نتيجة التقييم

Fig. 14. www.ush.sd.

0.37	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.2	السمعة
1	سهولة الاستخدام
0.3	الأمنية
0.56	النتيجة
جيد	نتيجة التقييم

Fig. 9. www.siu-sd.com.

0.3	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.2	السمعة
0.6	سهولة الاستخدام
0.3	الأمنية
0.43	النتيجة
ضعيف	نتيجة التقييم

Fig. 15. www.ribat.edu.sd.

0.37	النواحي الجمالية
0.9	الوسائط المتعددة
0.3	جودة المحتوى
0.5	السمعة
1	سهولة الاستخدام
0.7	الأمنية
0.7	النتيجة
جيد جدا	نتيجة التقييم

Fig. 10. www.bahri.edu.sd.

0.37	النواحي الجمالية
0.9	الوسائط المتعددة
0.5	جودة المحتوى
0.2	السمعة
0.6	سهولة الاستخدام
0.3	الأمنية
0.46	النتيجة
ضعيف	نتيجة التقييم

Fig. 11. www.iua.edu.sa.

The final evaluation of educational websites is calculated based on the following formula:

$$E.W=0.1*tot_aesthetic+0.3*tot_easeofuse+0.1*tot_multimedia+0.1*tot_richcontent+0.2*tot_reputation+0.2*tot_security. \quad (7)$$

- The scale of aesthetics, obtained from the sum of its sub-elements, takes the 10%.
- The measure of ease of use resulting from the sum of its sub-elements takes the 30%.
- The quality measure of the content generated from the sub-component takes a total of 10%.
- The resulting multimedia scale takes the 10% of its subgroups.
- The reputation scale obtained from the sum of its sub-elements is 20%.
- The resulting security measure takes 20% of the sub-component totals.

Table I depicts the Sudanese educational sites and its various evaluation results based on the proposed model. In this method, the authors took a sample of the Sudanese educational websites and evaluated them and put them in the form. The numbers that express the number of websites, for ease of presentation, the comparison and the analysis. This comparison is done after analyzing it through the following table. This will lead us to reach the various deficiencies in the educational sites in order to take care of them and then to address them.

TABLE I. SUDANESE EDUCATIONAL SITES AND EVALUATION RESULTS

Evaluation	Website
62	www.neelain.edu.sd
63	www.sustech.edu
43	www.uofk.edu
63	www.uofg.edu.sd
55	www.aau.edu.sd
56	www.siu-sd.com
70	www.bahri.edu.sd
46	www.iua.edu.sd
71	www.oiu.edu.sd
65	www.mu.edu.sd
62	www.ush.sd
43	www.ribat.edu.sd
85.25%	Average of Evaluation

VI. CONCLUSION AND FUTURE ENHANCEMENTS

The Internet has shown a rapid growth in terms of education which led to a new definition of almost all aspects of teaching. The Internet created a new teaching environment far different from anything that has come before. The explosion of the web has determined the need of measurement criteria to evaluate the aspects related to the quality of educational websites. Awareness of issues in quality issues has affected every websites in recent years, since an educational organization with a website is difficult to use and interact with gives a poor image on the Internet and weakens the position of it. Hence, it is important for a website especially the educational website to assess the quality of its e-service, in order to improve its quality. This paper proposes general criteria for evaluating the quality of Sudanese educational website in various evaluation criteria such as Aesthetics, Ease of Use, Rich Contents, Multimedia, Reputation and Security are depicted in detail long with its formulas. These criteria can be used by web site developers and its designers to create and to maintain the quality of educational websites do that the electronic service can be easily improved.

REFERENCES

[1] World Best Website Awards. (2007). Quality criteria for website excellence world best website awards. [online]. Available from <http://www.worldbest.com/criteria.htm> [2013, July 5].

[2] Khlaisang, J. (2010). Proposed Models of Appropriate Website and Courseware for E-Learning in Higher Education: Research Based

Design Models. Proceedings of the E-LEARN 2010 - World Conference on E-Learning in Corporate, Government, Healthcare & Higher Education organized by the Association for the Advancement of Computing in Education, Orlando, Florida, October 18-22, 2010.

[3] Khlaisang, J. (2012). Analysis of the Cultural Factors Affecting the Proper Design of Website and Electronic Courseware for e-Learning in ASEAN. Proceedings of the 26th Annual Conference of Asian Association of Open Universities (AAOU (, 2012 Chiba, Japan, October 16-18, 2012.

[4] Suliman Zakaria Suliman Abdall, Quality Assurance in Sudanese Higher Education: Current Status and Challenges Ahead, Journal of Total Quality Management, Vol 17, No. 1, 2016.

[5] Vera Silva Carlos, Ricardo Gouveia Rodrigues, Web site quality evaluation in Higher Education Institutions, Procedia Technology, Vol 5, pp.273-282, 2012.

[6] Michaelis, P., Ziesemer, T. Minimum quality standards and benchmarking in differentiated duopoly. JER (2020). <https://doi.org/10.1007/s42973-020-00050-y>.

[7] Jesse R. Sparks, Peter W. van Rijn & Paul Deane (2021) Assessing Source Evaluation Skills of Middle School Students Using Learning Progressions, Educational Assessment, 26:4, 213-240,

[8] Layla Hasan, Emad Abuelrub, Assessing the quality of web sites, Applied Computing and Informatics, Volume 9, Issue 1, pp. 11-29, 2011

[9] Sattler, C., & Sonntag, K. (2018). Quality cultures in higher education institutions—development of the Quality Culture Inventory. In P. Meusbürger, M. Heffernan, & L. Suarsana (Eds.), Geographies of the university (pp. 313–327). Cham: Springer International Publishing.

[10] Rahnuma, N. Evolution of quality culture in an HEI: critical insights from university staff in Bangladesh. Educ Asse Eval Acc 32, 53–81 (2020).

[11] Luis Ospina ,Daniela Godoy,Guillermo Lafuente Gustavo Rossi, “Assessing the quality of academic websites: a case study”, New review in Hypermedia and multimedia, pp 81-103, 2010.

[12] Zhang, P., Dran, G. 2001. Expectations and ranking of website quality features: results of two studies on user perceptions. In: Proceedings of the 34th Hawaii International Conference on System Sciences.

[13] Lautenbach, M.A.E., Schegget, I.S., Schoute, A.M., Witteman, C.L.M. 2006. Evaluating the Usability of Web Pages: A Case Study. Available at: <http://www.phil.uu.nl/preprints/ckipreprints/PREPRINTS/preprint011.pdf>.

[14] Yoo, S., Jin, J., 2004. Evaluation of the home page of the top 100 university web sites. Academy of Information and Management Sciences 8 (2), 57–69.

[15] Osborne, C., Rinalducci, J. 2002. Evaluation of Web Based Resources within the Art History Discipline, Technical Report, University of North Carolina.

[16] Singh, I., Sook, A. 2002. An evaluation of the usability of South African university web sites. In: Proceedings of the 2002 CITTE Conference, Durban, South Africa.

[17] Wang, E.S.T. (2016), “The moderating role of consumer characteristics in the relationship between website quality and perceived usefulness”, International Journal of Retail and Distribution Management, Vol. 44 No. 6, pp. 627-639.

[18] Wiranata, A.T. and Hananto, A. (2020), “Do website quality, fashion consciousness, and sales promotion increase impulse buying behavior of e-commerce buyers?”, Indonesian Journal of Business and Entrepreneurship, Vol. 6 No. 1, p. 74.

[19] Longstreet, P., Brooks, S., Featherman, M. and Loiacono, E. (2021), "Evaluating website quality: which decision criteria do consumers use to evaluate website quality?", Information Technology & People, Vol. ahead-of-print No. ahead-of-print. <https://doi.org/10.1108/ITP-05-2020-0328>.

[20] Loukas K. Tsironis, Educational websites quality assessment framework, International Journal of Decision Sciences, Risk and Management 2021 10:1-2, 51-77.

Secure Inter-Domain Routing for Resisting Unknown Attacker in Internet-of-Things

Bhavana A¹

Research Scholar, Department of Computer Science and
Engineering, VTU, Belagavi
Karnataka, India

Nanda Kumar A N²

Professor, Department of Computer Science and
Engineering, City Engineering College
Bangalore, India

Abstract—With an increasing adoption of Internet-of-Things (IoT) over massively connected device, there is a raising security concern. Review of existing security schemes in IoT shows that there is a significant trade-off due to non-adoption of inter-domain routing scheme over larger domain of heterogeneous nodes in an Internet of Things (IoT) via gateway nodes. Hence, the purpose of the proposed study is to bridge this trade-off by adopting a new security scheme that works over an inter-domain routing without any apriori information of an attacker. The goal of the proposed framework is to identify the malicious intention of attacker by evaluating their increasing attention over different types of hop links information. Upon identification, the framework also aims for resisting attacker node to participate in IoT environment by advertising the counterfeited route information with a target of misleading the attackers promoting autonomous self-isolation. The study outcome shows proposed scheme is secure compared to existing scheme.

Keywords—Internet-of-things; security; inter-domain routing; gateway node; attacker

I. INTRODUCTION

Internet-of-Things (IoT) is one of the evolving technologies toward data acquisition and controlling of large number of objects, digital and mechanical machines, computing devices, etc. without any form of dependencies toward human intervention [1]. Owing to the connection of different types of devices (or machines), there is an increasing security threats [2]. The first security threat in IoT device is inappropriate access control due to usage of same default password as the underlying firmware basically runs a default setting of all IoT devices of same model [3]. The second prominent threat is there is a larger base of attacker as different number of machines is connected over internet with an open port [4]. As IoT has an inclusion of large number of connected machines so eventually it suffers from regular updating of software. Conventional IoT device doesn't uses sophisticated encryption process and hence it gives rise to man-in-middle attack and denial-of-service attack in IoT [5]. Absence of reliable and trusted operating environment is another reason behind the security threat which often gives rise to privacy preservation issues. Due to usage of large number of machines, it is less feasible to offer reliable physical security towards IoT device. Apart from this, the conventional IoT nodes perform communication using various types of routing protocols [6] while Datagram Transport Layer Security (DTLS), Internet Protocol Security (IPSec), and Routing Protocol for Low-Power and Lossy Networks (RPL) are known to offer security.

However, there are different ranges of literatures which have reported of security pitfalls in existing routing protocols in IoT. Out of all this, one elementary concern is that IoT which runs on heterogeneous nodes doesn't seem to consider adopting inter-domain routing protocols. At present, there are such protocols reported to work over internet, however, they were never meant to be functional over IoT architecture, which is more complex form of architecture to be used in Future Internet Architecture [7][8]. Some of the challenges of implementing inter-domain routing scheme in IoT are as follows: i) developing a routing strategy among different forms of devices with multiple roles is definitely not an easy task considering the massiveness of the network, ii) developing both centralized as well as decentralized trusted authority connected to gateway node in IoT is one of the tedious task to be accomplished, iii) existing firewall system in IoT application is dependent on definition of patch and hence they are incapable of identifying new form of threats that are not defined in firewall system, iv) usage of conventional encryption process also comes with different forms of operational and communication challenges over resource constrained IoT devices [9]. All the above reasons serves as a motivation factor as well as reason towards develop a robust security protocol which is compliant of inter-domain routing as well as which is computationally efficient for practical implementation of complex environment of an IoT. The primary objective of the proposed manuscript is to introduce a novel solution where hop-based behaviour for all the nodes are observed to identify the malicious intention of attacker node, assuming its originality is unknown to the system. The secondary objective of proposed study is to resist attack in the form of novel inclusion of guard node. This new variant of node is meant to offer forged information of routes to attacker node in order to force them to accept the wrong direction of data dissemination. The objective of this operation is to ultimately results in either exclusion of attacker or their complete drainage of resources.

The organization of this manuscript is as follows: Section II discusses existing literatures of secure communication in IoT followed by discussion of research problems that are identified to be addressed in proposed study in Section III and proposed solution towards resisting unknown threat using inter-domain routing in IoT is briefed in Section IV. Section V discusses about algorithm design and implementation for secure route formulation and resisting malicious node participation followed by discussion of result analysis in Section VI. Finally, the conclusive remarks are provided in Section VII.

II. RELATED WORK

This section presents a briefing of the existing research implication being carried out towards securing communication in IoT as a continuation of our prior study [10]. The recent study carried out by Yilmiz et al. [11] have presented a discussion of a machine learning approach for securing IoT device using Routing Protocol for Low-Power and Lossy Network (RPL) protocol. Yazdinejad et al.[12] have used blockchain-based method for securing software defined network in IoT in the form of clustering. Study towards prevention of intrusion event is carried out by Haseeb et al. [13] considering the case study where sensors are used in IoT considering multi-hop routing and blockchain-based scheme. The work of Mick et al. [14] has presented a unique authentication scheme considering named data networking adhering to the concept of hierarchical routing. The work carried out by Xu et al.[15] have presented a secure routing scheme for resisting jamming attacks in IoT using game theory for exploring the optimal secure path for data delivery. Raof et al.[16] have presented discussion of existing threats and countermeasures exclusively towards frequently used RPL protocol in IoT. Haseeb et al. [17] have presented a security scheme where secret shares has been used for data communication with energy efficiency. Usage of reinforcement learning scheme has been noticed in work of Guo et al. [18] to ensure balance between security and quality of service at same time. Raof et al.[19] have presented an improved security scheme for RPL when subjected to different forms of attacks in IoT. The work carried out by Shin et al.[20] have developed an optimization mechanism for routing process in IoT focusing on securing authentication process. Wadhaj et al.[21] have developed a preventive technique towards attack on IoT device using RPL protocol with a target to maximize the reliability score of attacker identification process. Saleem et al.[22] have used a bio-inspired approach towards securing IoT communication over 5G. Ramos et al.[23] have carried out an investigation toward analyzing security aspects of resource-constrained IoT devices using probabilistic model. Liu et al. [24] have implemented a scheme towards resisting sink hole attack in IoT using probing routes considering consumption of network energy. Usage of geometric-based communication scheme anonymously is presented by Sun et al.[25] where hashing-based encryption has been utilized to ensure data privacy. Haseeb et al.[26] have presented a trust-based security scheme for mesh network in IoT considering cost of link and dissemination of data. Similar scheme has been carried out by Jhaveri et al.[27] towards trust-based security in IoT focusing on identifying the pattern of attack. Sathyadevan et al. [28] have introduced an authentication scheme using key generation technique exclusively meant for edge computing IoT device. Trust-based security scheme using provisioning approach was presented by Dass et al. [29] considering transport system in IoT. Agiollo et al. [30] have presented a unique scheme of identification of routing attack when standard RPL is deployed in IoT. Hence, there are different variants of security scheme toward safeguarding communication system in IoT. A closer look into all the above research implication has proven its substantial benefits from security perspective; however, they are highly symptomatic in nature of attack and is also

associated with various limitation. The next section outlines the identified research problem from the above stated literatures.

III. RESEARCH PROBLEM

The discussion of the research problem is carried out with respect to observed limitation and research gap as briefed below:

A. Limitation

The limitations that have been identified in proposed study are as follows:

- Existing security techniques towards IoT mainly uses either trust-based, or machine learning, or block chain in increasing pattern, which are sophisticated process for low resource IoT device.
- All the existing schemes has a well definition of attack and their strategy to initiate an attack is well known prior implementing security scheme.
- There are no reported study towards identification of threats on the basis of hops and malicious behaviour of attackers in heterogeneous nodes in IoT.
- There are no reported inter-domain routing scheme in IoT apart from the standard routing scheme which are exercised from long time.

B. Research Gap

The prime research gap of existing system is that with an increase of dynamicity and uncertainty in attack behaviour, existing security solutions over an IoT are yet not equipped to meet security demands both from hardware, software, and network perspective. The prime justification behind this research gap is that-it can be seen that there are various ranges of literatures that emphasize towards resisting attacks in IoT system, however, their work is not carried out over inter-domain routing system. This will be the prime reason that existing models are just theoretical model with theoretical proof of concept. The moment, such models are implemented over a gateway node, there is a need of a drastic revision towards such model in terms of network configuration as well as threat modelling. Hence, there is a vast gap between the security demands and the conclusive claims of existing studies.

Therefore, the problem statement of the proposed study can be stated as “Identifying an unknown attacker and resisting them in large IoT heterogeneous network using inter-domain routing scheme is quite a challenging task”. The next section discusses about the solution towards this problem.

IV. PROPOSED SYSTEM

The proposed study is a continuation of our prior framework of inter-domain routing with scalability [31] and interoperability [32] which offers a concrete baseline of two heterogeneous domains and wireless nodes within it to communicate via base station in IoT. The proposed system introduce security on the top of the previous framework for two purpose viz. i) to offer secure communication among communicating nodes and ii) to prevent any form of malicious nodes participating in data dissemination process. The

architecture of proposed system is as shown in Fig. 1. The core ideology of proposed secure inter-domain routing scheme is that every IoT device is communicated via a relay node controlled by base station which mainly broadcast hello message and instructions to control the topology. Hence, it becomes important for system to safeguard such relay nodes as well as other regular IoT nodes. The core operation is classified into secure route formulation and preventing attackers node to join the network on the basis of evaluation of links and control messages. A target node (exploited node) is assessed using

primary and secondary rule to find out if they are completely compromised or could have feasibility to be secured. Further, all the double hop links are evaluated in order to find out presence of malicious nodes. The novelty of proposed system is the formulation of guard node which is meant for preventing the malicious node from participating in data forwarding process. The next section of the paper elaborates about the algorithm design and implementation towards secure inter-domain routing in IoT.

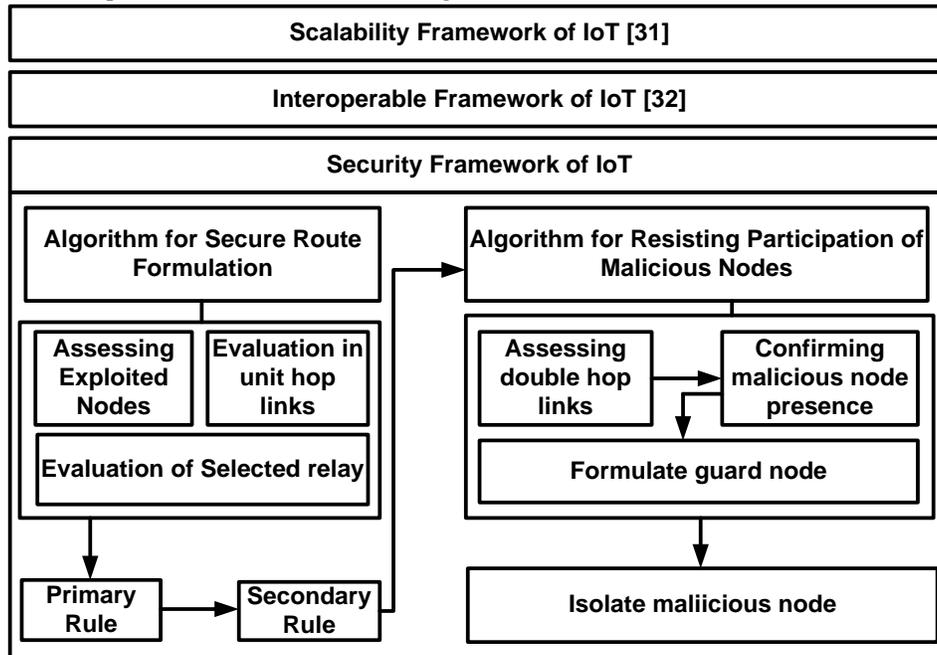


Fig. 1. Proposed Architecture of Secure Inter-domain Routing in IoT.

V. ALGORITHM IMPLEMENTATION

This section discusses about the algorithm design used for securing the proposed inter-domain routing system focusing on IoT use case [31] [32]. It should be noted that both these framework has already offered scalability as well as interoperability features while performing data transmission. This part of implementation focusses on embedding secure communication of inter-domain routing among all the participating nodes in IoT. The complete algorithm implementation is carried out using two discrete modules i.e. securing route formulation and resisting participation of malicious node. The description of algorithms design is as follow:

A. Algorithm for Secure Route Formulation

The main purpose of this algorithm is to initiate a secure topology in inter-domain routing connecting all sorts of nodes with an emphasis towards the exploited node (or compromised node). However, the degree of exploitation is yet not ascertained prior to implementation of this algorithm and the basis task of this algorithm is also to restrict all the communication system with unit hops in order to prevent collateral spread of exploitation by the unknown malicious node. The algorithmic flow is shown in Fig. 2 and its steps are as follows:

Algorithm for Secure Route Formulation

Input: n (wireless nodes)
Output: M (matrix storing network information)
Start
 1. **For** $i=1:n$
 2. $x_7 \rightarrow bc(uh(x_7))$
 3. Φ confirm $n_{dec}(x_7) \notin uh(\Phi)$
 4. **For** $j=1:\alpha$
 5. Φ assess $\beta \in uh(\alpha)$
 6. $\beta \notin bc(msg)$
 7. $\beta \rightarrow (uh+2)\Phi$
 8. **End**
 9. $M=[\Phi \beta uh]$
 10. **End**
End

The algorithm takes the input of all the participating wireless nodes n which after processing should yield a matrix M that stores network information to be used further for secure routing. The algorithm implements two set of rules to offer security. The primary rule is that the node x_7 will broadcast bc a unit hop links of x_7 node (Line-2). In such case, the exploited node Φ is required to confirm that declaration given by node x_7 should not belong to unit hop links of itself i.e. $uh(x_7)$ (Line-3).

This is possible by evaluating the previous broadcast message to assess if they have declared the transmitting node as its adjacent nodes. It is necessary that node x_7 must choose relay node in double hop $dh(x_7)$ in order to reach all the nodes present in double hop i.e. x_1 and x_4 . However, there is also a possibility that x_7 could opt for selecting Φ as its relay node in order to protect x_1 and x_4 nodes. Hence, according to security definition of non-repudiation, the node Φ is not permitted to deny the selection process. In such condition, the node Φ is incapable of confirming the fact if node x_7 is really an attacker node. However, it is feasible for the node Φ to assess if node x_7 has selected different relay node from the double hop links i.e. $dh(x_7)$ i.e. either x_2 or x_5 (Fig. 2). Therefore, a secondary ruleset is developed which states that if there is a presence of a different node α (Line-4) that is declared in the control message of inter-domain routing, than it is basic duty of the node Φ to find out if there is presence of some other new node say β which is already existing in unit hop links of α i.e. $uh(\alpha)$ (Line-5). It is also required to ensure that the node β is not declared in the transmitting message broadcasted (Line-6) as well as it is also required to ensure that this node β is positioned with a difference of three hops from the node Φ (Line-7). Once, this condition is evaluated, than the system undergoes another level of assessment which is to check if the node x_7 has selected some other new node which is a present in definition of unit hop links of node x_7 i.e. $uh(x_7)$ as relay node in order to protect other node β . All this information are stored in a matrix M (Line-9) which is consistently updated in every round of communication by the participating node. The core idea is to ensure that no unknown node is given the right to select some other undefined node from both the forms of link (unit/double) as the relay node.

The contribution of this algorithm are as following: i) a secure link is formulated among all the participating nodes, ii) multiple level of assessment is carried out to double-check the presence of relay node and its connection with all the adjacent nodes, iii) the algorithm is completely non-iterative and its information gets periodically updated in matrix M stating that there is a less computational complexity associated with it.

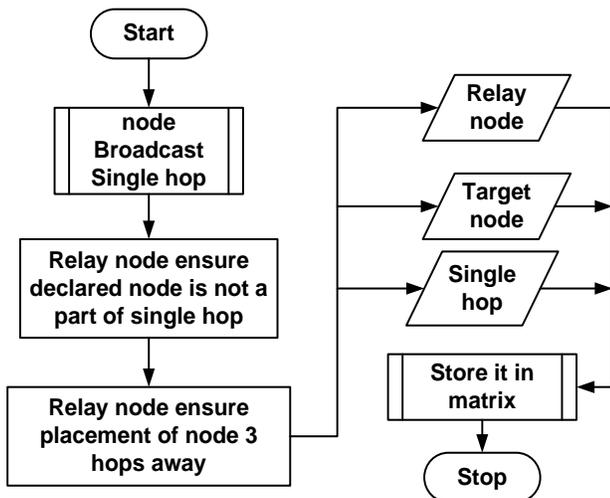


Fig. 2. Process Flow for Algorithm for Secure Route Formulation.

B. Algorithm for Resisting Participation of Malicious Nodes

This algorithm is basically responsible for confirming the presence of an unknown malicious nodes followed by a unique process of isolating them from the rest of the resured network formulated in matrix M . The core target of this algorithm is equivalent to first algorithm i.e. protecting the relay nodes from wrongly appointed by any attacker node. The prime concept underlying in this process is that attacker node is always curious to travel in double hop links in order to propagate their malicious code and the idea of this algorithm is to stop this process. The algorithmic flow (Fig. 4) and its respective steps are as follows:

Algorithm for Resisting Participation of Malicious Nodes

Input: Φ (exploited node), M (matrix of links)

Output: s_r (secure removal)

Start

1. **For** $i=1:\Phi$
2. **For** $cond=True$
3. Φ add n_g
4. Ensure $dist(\alpha, \beta) < (uh+2)|M$
5. $n_g \notin uh(\Phi)$
6. $\beta \rightarrow bc(n_g)$ & goto step-3
7. **Else**
8. remove n_g
9. flag $s_r \rightarrow$ secure removal of malicious node
10. **End**

End

The prime ideology of this algorithm are as follows: i) the node x_7 should demand to know only the nodes which is advertised by unit hop links of Φ i.e. $uh(\Phi)$, ii) the node x_7 selects relay node in order to achieve coverage to nodes mentioned in double hop links i.e. $dh(x_7)=x_1, x_4, x_2, x_5, x_8$ etc. It will mean that node x_7 is likely not to opt for x_4 node as its relay node can reach node x_2 via node x_5 as shown in Fig. 3.

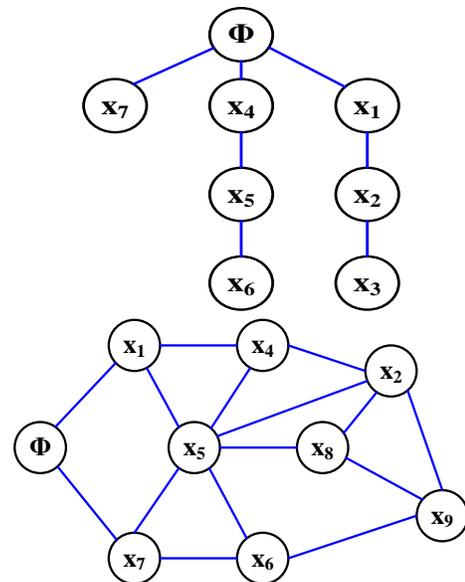


Fig. 3. Considered Test Topology.

The proposed algorithm formulates a condition $cond$ (Line-2) which means for all the nodes β that is a part of $dh(\Phi)$ if there exist any node α that is an element of unit hop of Φ i.e. $uh(\Phi)$ (Line-1). Therefore, the algorithm selects a guard node ng to be added for all node Φ (Line-3) such that spatial distance among all the nodes i.e. α, β is less than 3 hops (Line-4) obtained from matrix M .

The algorithm also ensure that this guard nodes ng is not advertised by unit hop links of node Φ i.e. $uh(\Phi)$ (Line-5) in order to protect them from getting disclosed to attacker node. The node β starts declaring guard node ng as a regular node in order to attract the attention of attacker (Line-6). In this case, as the attacker is also obeying the policy of undeniability of service in proposed secure inter-domain routing, therefore, it has to agree on accepting the counterfeited route information provided by β . This causes the attacker node to explore all the nodes which doesn't exist as well as which are never mentioned in either of the unit/double hop links of β or Φ or x_7 node. By following the counterfeited routes, the attacker allocates all its resources to capture information of the nodes and it drains all its resources until it either chooses to leave the network or stays in the network until its resources are completely drained. At the same time, it is also required to eliminate the guard node identity from the advertised message after the work of transmitting the counterfeited message is accomplished in order to offer more security. It also prevents the attacker even to guess the formation as well as trend of guard node message especially in case of multiple attackers. Finally, a flag message of secured removal of malicious node is disseminated in the network reporting the identity of the attacker node that prevents the same attacker node to intrude the network. The next section discusses about result being obtained.

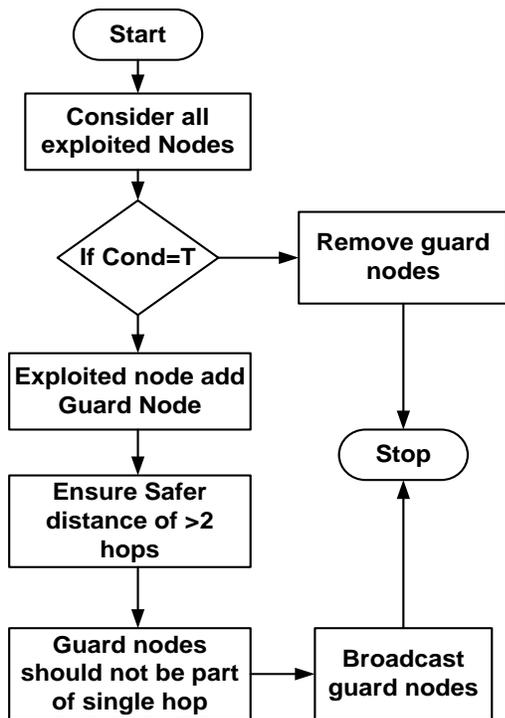


Fig. 4. Process Flow for Resisting Participation of Attacker.

VI. RESULT ANALYSIS

This section discusses about the results obtained after implementing the proposed algorithm discussed in prior section. A simulation area of 1000×1000 m² is used where 100 sample wireless nodes are deployed adhering to the inter-domain routing scheme [32]. The proposed logic is scripted in MATLAB where different test environment of undefined attacker is considered. The outcome of the study has been evaluated by different parameters. Table I and Fig. 5 highlights the frequencies of an attack event for 800 node density that clearly highlights the reduction of attack event with progressive density of nodes. The justification behind this outcome is that with more events of positively identified attacks, the routing tables gets updated which can be accessed via any gateway node to upgrade heterogenous domains under communication. Hence, the proposed system offers better control of malicious nodes in inter-domain routing scheme.

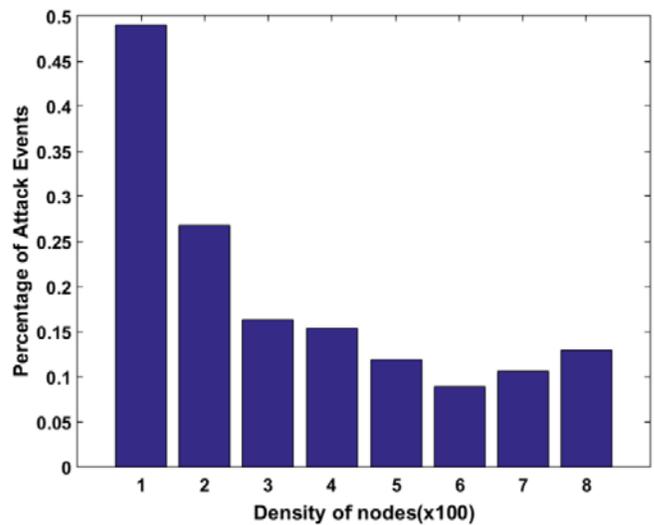


Fig. 5. Percentage of Attack Event.

TABLE I. NUMERICAL SCORE OF ATTACK EVENTS

Density of Nodes	Percentage of Attack Events
1	0.489
2	0.275
3	0.17
4	0.15
5	0.13
6	0.09
7	0.11
8	0.14

The proposed system has been compared with existing security protocol in IoT Routing Protocol for Low-Power and Lossy Networks (RPL), which is claimed to offer balance between security and resource efficiency. The idea is to assess the control towards overhead as well as dependencies of guard nodes. Fig. 4 highlights that proposed system offers reduced overhead that is computed by every extra data being forwarded by the transmitting node. It is because although RPL offers great security but it suffers from long delays especially when

exposed to unknown form of attacks under node formation in tree. However, proposed system performs parallel confirmation of node legitimacy as well as data transmission causing reduced overhead.

Discussion: From the tabulated information as well as graphical data, it can be seen that proposed system is potential enough to control the attacker (Fig. 5) as well as it can also reduce the overhead (Fig. 6 and Table II). The significance of this outcome is quite high as usage of conventional scheme of IoT secure routing results in increasing overhead. When subjected to inclusion of multiple hardware in the form of network devices, it is quite inevitable that IoT device will incur more number of queued packets resulting in overhead. However, this is not the case with proposed scheme for two reason viz. i) all the hop information are basically shared among all the regular nodes and hence accessibility becomes easier, and ii) permission for data transmission is granted only after a node is confirmed to be a legitimate node in progressive round.

Fig. 7 and Table III highlights that proposed system offers reduced dependencies of guard node in order to prevent the malicious node as compared to existing protocol of RPL. Fig. 8 highlights the comparative analysis of processing time which shows that proposed system consumes much less time in contrast to existing RPL protocol.

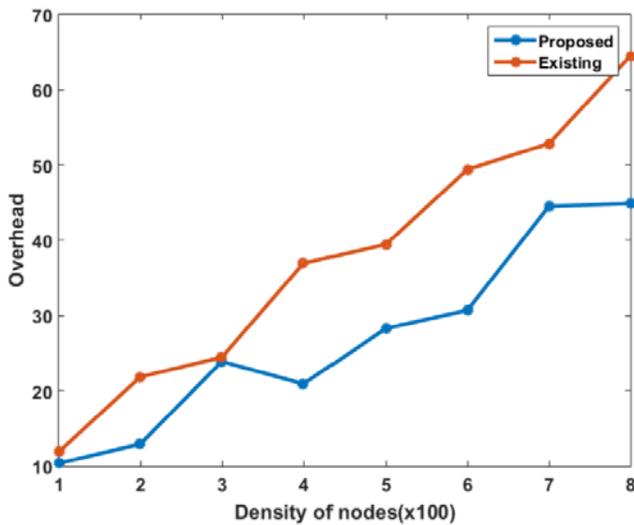


Fig. 6. Comparative Analysis of Overhead.

TABLE II. NUMERICAL SCORE FOR OVERHEAD ANALYSIS

Density of Nodes	Existing System	Proposed System
1	10.38	11.93
2	12.96	21.87
3	23.91	24.48
4	20.94	36.96
5	28.31	39.47
6	30.71	49.42
7	44.53	52.87
8	44.90	64.49

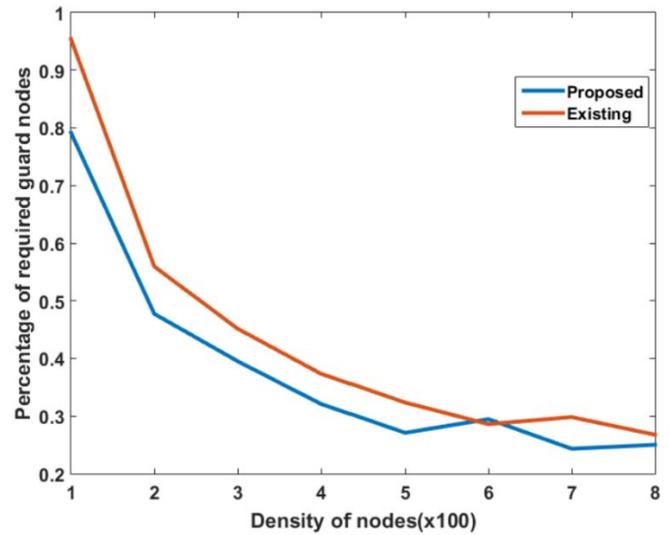


Fig. 7. Comparative Analysis of Percentage of Required Guard Nodes.

TABLE III. NUMERICAL SCORE FOR GUARD NODE DEPENDENCY

Density of Nodes	Existing System	Proposed System
1	0.7913	0.9539
2	0.4774	0.5594
3	0.3952	0.4516
4	0.3211	0.3735
5	0.271	0.3235
6	0.2947	0.286
7	0.2434	0.2983
8	0.2502	0.2677

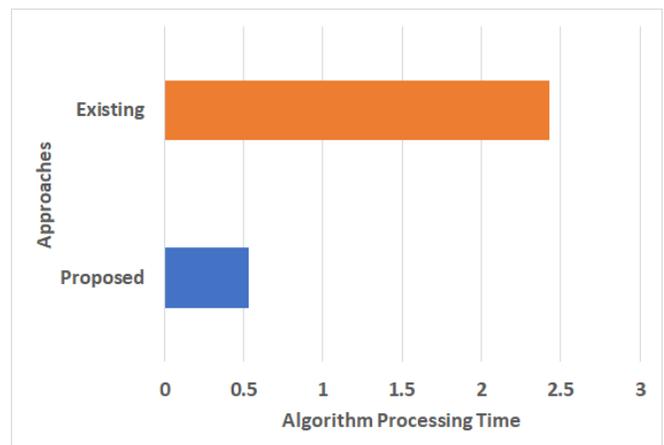


Fig. 8. Comparative Analysis of Algorithm Processing Time.

Discussion: The prime reason behind this outcome shown in Fig. 7 and Fig. 8 is as follows- because RPL protocol performs assessment of security instantly and performs secure encryption ignoring the fact that if the same attacker has compromised some other set of links in its far neighboring nodes. On the other hand, proposed system gather more information about attacker node via guard node and this gets updated in the form of matrix, which is easily accessible via gateway node. This causes the attacker node to completely

eliminate from network by spending all its resources towards counterfeited nodes advertised by guard node and if the same attacker or its connected attacker with multiple strategy is trying to launch an attack from different nodes. The proposed system easily captures that information via double hop links. Greedy attackers have more concentration of response towards double hop links and that makes the identification quite easier. It is seen that RPL completely works on directed acyclic graph without any edges outgoing. Apart from this, owing to inclusion of number of control messages used, there is a huge consumption of time especially when working on higher number of heterogeneous nodes. This causes much consumption of its time, whereas proposed system formulates a simplified logic of capturing the attacker intention via their response message over the dual hop links. Identification operation becomes much easier by accessing a single hand matrix for faster detection. Hence, proposed system can be considered almost instantaneous in offering its response time, which is an additional benefit from secure routing.

VII. CONCLUSION

This paper has presented a unique solution towards confirming the malicious intention of an attacker over proposed secured inter-domain routing in IoT. The summary of research findings are as follows: i) one prime indicator of an attacker node is to assess their intention to carry out routing from the nodes with maximum hop, ii) trust computation always works well when it is splitted to local trust and global reputation system, iii) acceptance of global reputation system should be followed by authenticating the legitimacy of the neighboring nodes, iv) updating hop table as well as limiting hop access is one of the safest means to restrict the propagation of uncertain threats. The summary of the proposed method are i) proposed method is capable of working over an inter-domain routing in presence of uncertain threat, ii) proposed model exploits the hop-based detailed information to formulate the attack possibilities as well as malicious intention, iii) proposed model offers robust security even without using conventional encryption process in IoT. The summary of contribution of the proposed system are i) it presents a novel architecture where secure inter-domain routing is implemented for resisting unknown attacker, ii) the complete analysis of malicious intention is based on attacker response towards different types of hops, iii) the framework also present an inclusion of a guard node which is meant for forwarding forged routing information to mislead the attacker node. The novelty of the proposed study are as follows: i) the model is independent of any form of attack definition unlike existing system which demands proper definition and types of attack, ii) a novel selection of relay node is developed unlike any secure routing scheme in IoT for topology control, iii) a completely non-encryption-based approach whereas majority of standard approaches uses cryptography.

REFERENCES

- [1] H. H. Qasim, A. E. Hamza, L. Audah, H. H. Ibrahim, H. A. Saeed, M. I. Hamzah, "Design and implementation home security system and monitoring by using wireless sensor networks WSN/internet of things IoT", International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 3, June 2020, pp. 2617~2624.
- [2] Shamshekhar S. Patil, Arun Biradar, "Novel authentication framework for securing communication in internet-of-things", International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 1, February 2020, pp. 1092~1100.
- [3] Mohammed Al- Shabi, Anmar Fakhri Abuhamdah, "Using deep learning to detecting abnormal behavior in IoT", vol.12, No.2, 2022, DOI: <http://doi.org/10.11591/ijece.v12i2.pp%25p>.
- [4] Basheer Al-Duwairi, Wafaa Al-Kahla, Mhd Ammar AlRefai, Yazid Abdelqader, Abdullah Rawash, Rana Fahmawi, "SIEM-based detection and mitigation of IoT-botnet DDoS attacks", International Journal of Electrical and Computer Engineering (IJECE), Vol. 10, No. 2, April 2020, pp. 2182_2191.
- [5] Azka Wani, S. Revathi, "Ransomware protection in IoT using software defined networking", International Journal of Electrical and Computer Engineering (IJECE) Vol. 10, No. 3, June 2020, pp. 3166~3175.
- [6] H. Kim, J. Ko, D. E. Culler and J. Paek, "Challenging the IPv6 Routing Protocol for Low-Power and Lossy Networks (RPL): A Survey," in IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2502-2525, Fourthquarter 2017, doi: 10.1109/COMST.2017.2751617.
- [7] A. M. Alberti, G. D. Scarpioni, V. J. Magalhães, A. Cerqueira S., J. J. P. C. Rodrigues and R. da Rosa Righi, "Advancing NovaGenesis Architecture Towards Future Internet of Things," in IEEE Internet of Things Journal, vol. 6, no. 1, pp. 215-229, Feb. 2019, doi: 10.1109/JIOT.2017.2723953.
- [8] T. M. Fernández-Caramés, "From Pre-Quantum to Post-Quantum IoT Security: A Survey on Quantum-Resistant Cryptosystems for the Internet of Things," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6457-6480, July 2020, doi: 10.1109/JIOT.2019.2958788.
- [9] N. M. Karie, N. M. Sahri, W. Yang, C. Valli and V. R. KEBANDE, "A Review of Security Standards and Frameworks for IoT-Based Smart Environments," in IEEE Access, vol. 9, pp. 121975-121995, 2021, doi: 10.1109/ACCESS.2021.3109886.
- [10] Bhavana A, "Evaluating Perception, Characteristics and Research Directions for Internet of Things (IoT): An Investigational Survey", International Journal of Computer Applications (0975 – 8887) ,Volume 121 – No.4, July 2015.
- [11] S. Yilmaz, E. Aydogan and S. Sen, "A Transfer Learning Approach for Securing Resource-Constrained IoT Devices," in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 4405-4418, 2021, doi: 10.1109/TIFS.2021.3096029.
- [12] A. Yazdinejad, R. M. Parizi, A. Dehghantanha, Q. Zhang and K. -K. R. Choo, "An Energy-Efficient SDN Controller Architecture for IoT Networks With Blockchain-Based Security," in IEEE Transactions on Services Computing, vol. 13, no. 4, pp. 625-638, 1 July-Aug. 2020, doi: 10.1109/TSC.2020.2966970.
- [13] K. Haseeb, N. Islam, A. Almogren and I. Ud Din, "Intrusion Prevention Framework for Secure Routing in WSN-Based Mobile Internet of Things," in IEEE Access, vol. 7, pp. 185496-185505, 2019, doi: 10.1109/ACCESS.2019.2960633.
- [14] T. Mick, R. Tourani and S. Misra, "LASER: Lightweight Authentication and Secured Routing for NDN IoT in Smart Cities," in IEEE Internet of Things Journal, vol. 5, no. 2, pp. 755-764, April 2018, doi: 10.1109/JIOT.2017.2725238.
- [15] Y. Xu, J. Liu, Y. Shen, J. Liu, X. Jiang and T. Taleb, "Incentive Jamming-Based Secure Routing in Decentralized Internet of Things," in IEEE Internet of Things Journal, vol. 8, no. 4, pp. 3000-3013, 15 Feb.15, 2021, doi: 10.1109/JIOT.2020.3025151.
- [16] A. Raouf, A. Matrawy and C. Lung, "Routing Attacks and Mitigation Methods for RPL-Based Internet of Things," in IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1582-1606, Secondquarter 2019, doi: 10.1109/COMST.2018.2885894.
- [17] K. Haseeb, N. Islam, A. Almogren, I. Ud Din, H. N. Almajed and N. Guizani, "Secret Sharing-Based Energy-Aware and Multi-Hop Routing Protocol for IoT Based WSNs," in IEEE Access, vol. 7, pp. 79980-79988, 2019, doi: 10.1109/ACCESS.2019.2922971.
- [18] X. Guo, H. Lin, Z. Li and M. Peng, "Deep-Reinforcement-Learning-Based QoS-Aware Secure Routing for SDN-IoT," in IEEE Internet of Things Journal, vol. 7, no. 7, pp. 6242-6251, July 2020, doi: 10.1109/JIOT.2019.2960033.
- [19] A. Raouf, A. Matrawy and C. -H. Lung, "Enhancing Routing Security in IoT: Performance Evaluation of RPL's Secure Mode Under Attacks," in

- IEEE Internet of Things Journal, vol. 7, no. 12, pp. 11536-11546, Dec. 2020, doi: 10.1109/JIOT.2020.3022276.
- [20] D. Shin, K. Yun, J. Kim, P. V. Astillo, J. Kim and I. You, "A Security Protocol for Route Optimization in DMM-Based Smart Home IoT Networks," in IEEE Access, vol. 7, pp. 142531-142550, 2019, doi: 10.1109/ACCESS.2019.2943929.
- [21] I. Wadhaj, B. Ghaleb, C. Thomson, A. Al-Dubai and W. J. Buchanan, "Mitigation Mechanisms Against the DAO Attack on the Routing Protocol for Low Power and Lossy Networks (RPL)," in IEEE Access, vol. 8, pp. 43665-43675, 2020, doi: 10.1109/ACCESS.2020.2977476.
- [22] K. Saleem, G. M. Alabduljabbar, N. Alrowais, J. Al-Muhtadi, M. Imran and J. J. P. C. Rodrigues, "Bio-Inspired Network Security for 5G-Enabled IoT Applications," in IEEE Access, vol. 8, pp. 229152-229160, 2020, doi: 10.1109/ACCESS.2020.3046325.
- [23] A. Ramos, R. T. P. Milfont, R. H. Filho and J. J. P. C. Rodrigues, "Enabling Online Quantitative Security Analysis in 6LoWPAN Networks," in IEEE Internet of Things Journal, vol. 6, no. 3, pp. 5631-5638, June 2019, doi: 10.1109/JIOT.2019.2904302.
- [24] Y. Liu, M. Ma, X. Liu, N. N. Xiong, A. Liu and Y. Zhu, "Design and Analysis of Probing Route to Defense Sink-Hole Attacks for Internet of Things Security," in IEEE Transactions on Network Science and Engineering, vol. 7, no. 1, pp. 356-372, 1 Jan.-March 2020, doi: 10.1109/TNSE.2018.2881152.
- [25] Y. Sun et al., "Lightweight Anonymous Geometric Routing for Internet of Things," in IEEE Access, vol. 7, pp. 29754-29762, 2019, doi: 10.1109/ACCESS.2019.2902621.
- [26] K. Haseeb, I. Ud Din, A. Almogren, N. Islam and A. Altameem, "RTS: A Robust and Trusted Scheme for IoT-Based Mobile Wireless Mesh Networks," in IEEE Access, vol. 8, pp. 68379-68390, 2020, doi: 10.1109/ACCESS.2020.2985851.
- [27] R. H. Jhaveri, N. M. Patel, Y. Zhong and A. K. Sangaiah, "Sensitivity Analysis of an Attack-Pattern Discovery Based Trusted Routing Scheme for Mobile Ad-Hoc Networks in Industrial IoT," in IEEE Access, vol. 6, pp. 20085-20103, 2018, doi: 10.1109/ACCESS.2018.2822945.
- [28] S. Sathyadevan, K. Achuthan, R. Doss and L. Pan, "Protean Authentication Scheme – A Time-Bound Dynamic KeyGen Authentication Technique for IoT Edge Nodes in Outdoor Deployments," in IEEE Access, vol. 7, pp. 92419-92435, 2019, doi: 10.1109/ACCESS.2019.2927818.
- [29] P. Dass, S. Misra and C. Roy, "T-Safe: Trustworthy Service Provisioning for IoT-Based Intelligent Transport Systems," in IEEE Transactions on Vehicular Technology, vol. 69, no. 9, pp. 9509-9517, Sept. 2020, doi: 10.1109/TVT.2020.3004047.
- [30] A. Agiollo, M. Conti, P. Kaliyar, T. -N. Lin and L. Pajola, "DETONAR: Detection of Routing Attacks in RPL-Based IoT," in IEEE Transactions on Network and Service Management, vol. 18, no. 2, pp. 1178-1190, June 2021, doi: 10.1109/TNSM.2021.3075496.
- [31] A. Bhavana, A. N. Nandha Kumar, "An Analytical Modeling for Leveraging Scalable Communication in IoT for Inter-Domain Routing", Springer-Proceedings of the Computational Methods in Systems and Software, pp.1-11, 2018.
- [32] A. Bhavana, A. N. Nandha Kumar, "ICS: Interoperable Communication System for Inter-Domain Routing in Internet-of-Things", SAI-The Science and Information Organization, vol.10, Iss.5, 2021, 10.14569/IJACSA.2021.0120533.

Snowball Framework for Web Service Composition in SOA Applications

Mohamed Elkholy¹

Computer Engineering Department
Pharos University in Alexandria
Alexandria, Egypt

Youcef Bagdadi²

Department of Computer Science
Sultan Qaboos University
Muscat, Oman

Marwa Marzouk³

Information Technology Department
Matroh University
Matroh, Egypt

Abstract—Service Oriented Architecture (SOA) has emerged as a promising architectural style that provides software applications with high level of flexibility and reusability. However, in several cases where legacy software components are wrapped to be used as web services the final solution does not completely satisfy the SOA aims of flexibility and reusability. The literature review and the industrial applications show that SOA lacks a formal definition and measurement for optimal granularity of web services. Indeed, wrapping several business functionalities as a coarse-grained web services lacks reusability and flexibility. On the other hand, a huge number of fine-grained web services results in a high coupling between services and large size messages transferred over the Internet. The main research question still concerns with “How to determine an optimal level of service granularity when wrapping business functionalities as web services?” This research proposes the Snowball framework as a promising approach to integrate and compose web services. The framework is made up three-step process. The process uses the rules in deciding the web services that have an optimal granularity that maintains the required performance. To demonstrate and evaluate the framework, we realized a car insurance application that was already implemented by a traditional approach. The results show the efficiency of snowball framework over other approaches.

Keywords—Service oriented architecture (SOA); web service granularity; web service composition; software flexibility; snowball composition framework

I. INTRODUCTION

SOA allows software systems to be composed as a group of loosely coupled software components called services [1]. SOA aims to provide cost effective flexible solution to business organizations [2, 3]. However, SOA had not gained an extreme popularity until the emerging of web service technology in early 2000s [4]. Since that time, web service became the main trend to implement SOA systems [5]. Several organizations tend to wrap legacy software components in the form of web services to implement SOA-based applications [6]. Wrapping legacy software into web services reduces the cost of implementing new software systems. However, in several cases where legacy components are wrapped to be (re)used as web services, the final solution does not completely satisfy the SOA aims of flexibility and reusability. The reason behind that is the unsuitable level of service granularity. Service granularity has two different perspectives: business perspective and IT perspective. From a business perspective, service granularity is associated with the amount of business tasks fulfilled with that

service. On the other hand from IT perspective, web service granularity is associated with size of data transferred from or towards the service as well as its code length [7].

Service granularity affects reusability, efficiency and performance of the services. Wrapping several business functionalities as a coarse-grained web services leads to a single use service [8]. Such service lacks reusability and flexibility since the separation of concerns and cohesion are missing. On the other hand, composing business tasks from large number of small fine-grained services leads to high coupling between services. Such situation leads to communication complexity and degraded performance. That is, an incorrect service granularity leads to bad performance, low reuse possibilities, inappropriate abstraction levels, and services without business value [9].

It is critical to balance between coarse-grained and fine-grained web services while mapping SOA design to individual web services [10]. Unfortunately, the literature lacks detailed studies about service granularity and its impact on reusability, flexibility, and performance [11].

Consequently, one of the main problems that faces developers while developing web services-based SOA is the difficulty to determine optimal service granularity, especially as there is no theoretical definition for service granularity in the literature.

The main research question still concerns with “How to determine an optimal level of service granularity when wrapping business functionalities as web services?”

This research proposes the Snowball as a promising approach to compose web services in SOA-based applications. Snowball is framework made up of a set of rules and a three-step process. The process uses the rules to check the right and optimal granularity of the services. It first decomposes a Business Process (BP) into smaller sub-processes that are further decomposed into business tasks, each of which is a set of activities. Next, it maps the tasks into individual fine-grained web services. Then it checks the fine-grained web services against the rules, in order to allow their integration. Finally, it optimizes the granularity.

Snowball aims at providing web services that have the optimal granularity while maintaining the required flexibility, reusability, and high performance in terms of low size of data transferred. It is meant to be used by organizations that want to

offer its functionalities to users as web services, and can also be used by organizations to build up their own business applications.

To demonstrate and evaluate the framework, we realized a car insurance application that was already implemented by a traditional approach. The results show the efficiency of snowball framework other traditional approaches.

Moreover, the proposed framework has an advantage over other composition frameworks that generally use Business Process Execution Language (BPEL). It integrates and composes services functionalities before the implementation phase. Hence, the framework allows three different modes of services: wrapping legacy components, invocation from a service provider, or creation from scratch (coding). Therefore, Snowball eliminates the utilization of glue code languages such as BPEL, which leads to degraded performance and hard validation tests.

This paper is organized as follows: Section 2 presents related work. Section 3 develops the Snowball framework, i.e., rules and methodology. Section 4 presents the results of the empirical study. The conclusion section summarizes the contribution, its limitation, its impacts, and future work.

II. RELATED WORK

One of the main challenges in web service applications concerns with the granularity of services. This section analyzes different popular methodologies for SOA applications, with respect to the granularity of web services. Several models tried to formalize different processes for an organization to adopt SOA [12]. However, fewer researches focused on services granularity and size of service messages.

SOMA is a popular SOA design framework, introduced by IBM, that models business functionality as coherent individual services. To implement new software for an organization, SOMA defines a domain decomposition approach to perform the design phase. The main idea is to decompose the business into logical coherent functional areas. Each area consists of related processes that are further split to smaller sub processes [13]. Each sub process is decomposed into a set of activities which are listed together to form service portfolio [14]. Each service's functionality in the service portfolio is assigned to a web service. SOMA has no restriction on service granularity or on the size of service messages, whether the service is from the legacy system or from external services. Hence, several SOMA designs that lead to large services that perform several individual functionalities, hence, missing the required flexibility [15]. Such situation leads to a set of non-reusable services neglecting SOA aims of software reusability [16]. On the other hand, SOMA application might be implemented with extremely high number of fine-grained services. Such implementation may lead to large size of data transfer while aggregating these services together [17].

Another popular model for SOA adoption is Service Oriented Architecture Maturity Model (SOAMM). SOAMM defines a model for monitoring different levels of development, implementation, and usage of SOA [18]. SOAMM defines a set of characteristics for organizational architecture that are essential for any organization to be able to implement web

services-based SOA. SOAMM defines service selection and collaboration between services from the business point of view only [19]. However, SOAMM does not define rules for service granularity from IT perspective such as size of input/output messages.

Thomas Erl [20] defined Mainstream SOA Methodology (MSOAM) as a framework to design, implement, test, and deploy web services. MSOAM identifies seven activities during analyses and design phases. It starts by Ontology definition, then perform business model Alignment. Further it performs service oriented design to develop services that fulfill each process of business functionality. This framework has an advantage in defining dependencies between services. However, it does not define how these dependencies can affect service granularity. Thus several MSOAM applications suffer from coarse-grained web services lacking flexibility and reusability.

Business Process-driven Methods [21] is considered one of the most common strategy used to identify services in SOA. This method uses clustering algorithm to identify services from the business perspective. Business elements are divided into rules and requirements, and then a syntax analysis is applied to perform service selection for each BP [22]. Such method focuses on BPs, and gives less attention to data transfer. The main drawback of this method is the extremely fine-grained services that lead to large amount of communication overheads between services. Implementing web service application with large number of fine grained services increases the size of messages required for services communication [23, 24].

This figures out the problems associated with service granularity while implementing SOA by using web services. The literature lacks theoretical methods to define optimal service granularity. Unsuitable level of service granularity leads to significant drawbacks in flexibility, efficiency and performance of SOA based applications [25]. The proposed framework assists developers in deciding the optimal granularity of web services that maintains flexibility and high performance.

III. PROPOSED SNOWBALL FRAMEWORK

The proposed snowball framework provides a systematic approach to determine the optimal service granularity for web services-based SOA, in terms of performance and efficiency.

It defines a set of rules and a three-step process. The rules specify mapping business tasks to IT web services. The rules also define the conditions under which two services or more should be integrated together. The three processes define the actions taken to apply the rules to the business tasks step by step till getting a suitable level of service granularity. It aims at assuring an optimal service granularity that satisfies lower coupling and higher cohesion.

A. Service Granularity

The framework considers two different properties of service granularity: (1) the business functional granularity, representing the number of elementary business tasks fulfilled by the service, and (2) the data granularity, concerning with size of input/output data included in the service messaging.

From the business perspective, the fine-grained service is the service that performs an atomic task [23]. While from IT perspective, a fine-grained service is the service that has a limited size of data transfer. Thus a service could be fine-grained from a business perspective, and coarse-grained from a data granularity perspective. For instance, a service that displays a map performs a single business task but carries a huge size of data.

B. Snowball Rules to Optimize Service Granularity

The framework provides two sets of rules to optimally and efficiently integrate fine-grained services together, considering both functional granularity and data granularity. Dependencies between services are also a point of concern.

1) *Rules to map business tasks into IT services:* Mapping business tasks into IT services consists in assigning each single business task to an elementary web service, i.e., an elementary coherent fine-grained service.

a) *Rule 1:* If a legacy software component satisfies a single business task, then it is wrapped to act as a web service with only one single operation.

b) *Rule 2:* If the required functionality exists in public/private registries as a web service with one operation, then select it.

c) *Rule 3:* If the service is to be locally implemented (by coding), then the code includes only one single operation.

Applying these rules results in a high flexibility and reusability of mapped web services. However, increasing the number of individual web services in an application affects its complexity and performance in terms of response time and large size of network traffic [26, 27]. Therefore, there is a need to integrate and compose service into an optimal granularity by using the following rules.

2) *Rules to integrate IT services:* After assigning elementary business tasks to IT service, the output is a set of fine-grained service. The following rules are applied to these services to achieve optimal granularity.

a) *Rule 1:* Two services S_i and S_j are integrated together if:

- The business workflow requires execution of the two services sequentially.
- The input parameters for S_i are the same as S_j or the output of S_i is the required input for S_j .

b) *Rule 2:* Two services S_i and S_j are integrated together if:

- S_i and S_j are in the same business domain and are connected to the same database tables and.
- S_i and S_j should be at the same branch of business workflow.

3) Factors that manages services integration:

Factor 1: S_i and S_j have sequential execution.

Factor 2: The output of S_i is an input for S_j .

Factor 3: S_i and S_j have the same I/O.

Factor 4: S_i and S_j have connection to the same database.

Factor 5: S_i and S_j have data dependencies, i.e., S_j cannot be executed until S_i is completed as S_j has one (not all) of its inputs passed from outputs of S_i .

C. Snowball Steps to Optimize Service Granularity

The Snowball process, shown in Fig. 1, consists of three main steps that should be completed to provide SOA applications with the required flexibility, reusability, and high performance of the services that compose them. Step 1 identifies business tasks, step 2 maps each business task into an IT service, whereas step 3 optimizes the service integration.

1) *Service identification:* Each BP is broken down into smaller sub-processes and then to single elementary tasks. An elementary task performs atomic coherent business functionality. Then all the elementary tasks are listed in a task table, as exemplified in Table I.

2) *Mapping business tasks into IT services:* This step uses the first aforementioned set of three rules to map business tasks into IT services. Mapping business tasks to a web service means selecting a web service that fulfills the business functionality of the task. Each atomic business task is mapped to an elementary web service that would be wrapped from the legacy systems or discovered over web service registries, or even implemented as a new web service. Different activities of mapping atomic business functionality to web services are shown in Fig. 2.

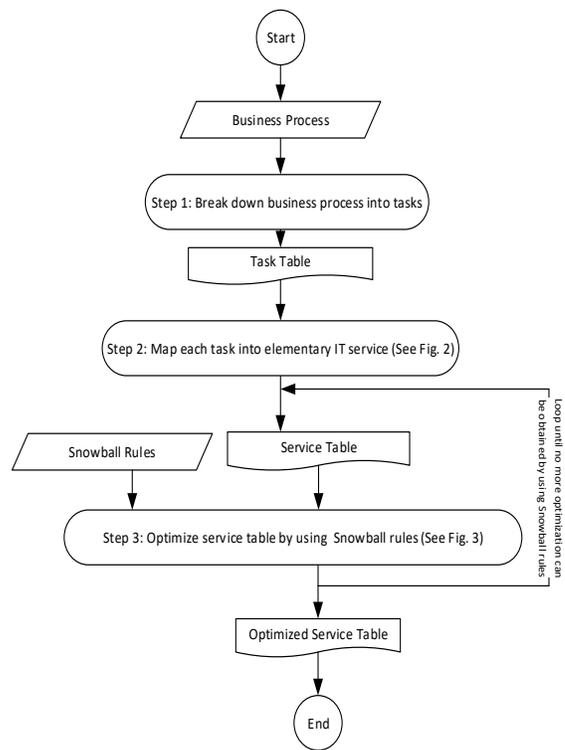


Fig. 1. Snowball Process.

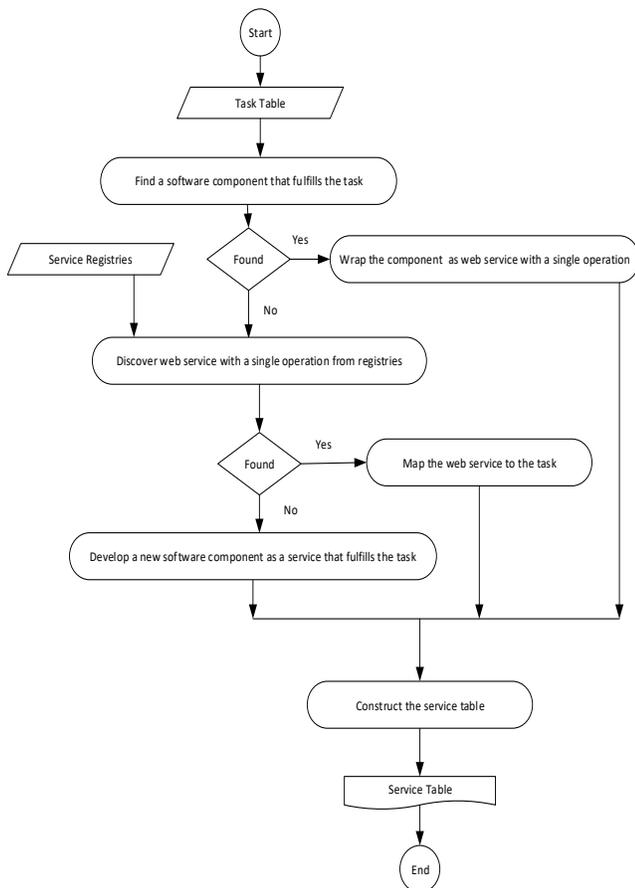


Fig. 2. Details of Step 2 of the Snowball Process.

Up till this step the system is composed of fine-grained IT services that ensure a high level of cohesion. All the atomic services are then listed in a services table. The services table lists all the used services, including their business functionality, I/O parameters, connected database, and dependent services. The dependency between services is divided into control flow dependency and dataflow dependency. The I/O parameters and connected database are chosen as they have the most effect on services coupling. The dependency between services is further used to construct the criteria of integration between two services or more.

3) *Optimization of the service table*: The third step is responsible of optimizing the integration of different individual services together. Such integration avoids building applications from fine-grained web services that increases the interaction between the application and outer invoked services. This scenario leads to poor performance. The integration process would reduce the total size of I/O messaging of the client application to maintain high performance. Unlike traditional composition such as BPEL, the integration in the Snowball process consists of adding the business functionality of the first service to the functionality of second service. Thus, the integration of two services functionalities results in new service that performs the functionalities of both services. Snowball integrates services with each other in recursive rounds. In each round, a service is

added to the existing one(s) to constitute a new integrated service. The newly integrated services will act as elementary services in the next round and so on till we get a service where no more integration process can be applied according to the rules. After each round the service table is updated to list the newly integrated services with their parameters. Hence, services integration process is repeated in recursive order by applying the second aforementioned set rules, in order to control service integration, as shown in Fig. 3.

In this step, Snowball framework applies the aforementioned second set of rules in order to achieve two main objectives for services integration regarding size of transferred data and database connections.

First objective: Minimize overall service interface messages.

This objective is achieved by applying rule1 regarding both the business workflow and the size of transferred data. If S1 and S2 have the same input parameters, they are integrated together in a new service S3. During runtime S3 will be invoked with the input parameters of S1 and returns the output parameters of S2 rather than sending the same data twice from S1 and S2. The decrease in the sent and received data between the application and outer web services has a great effect on performance, especially for huge size of parameters.

Second objective: Minimize Database connections.

This objective regards the connection between web services and databases. Rule 2 may not be available for discovered services as the Web Service Describing Language (WSDL) file almost contains the input/output parameters without information about database connections. However, if the web services are wrapped from the legacy software asset or implemented as new services, information about connections between data bases and services are available.

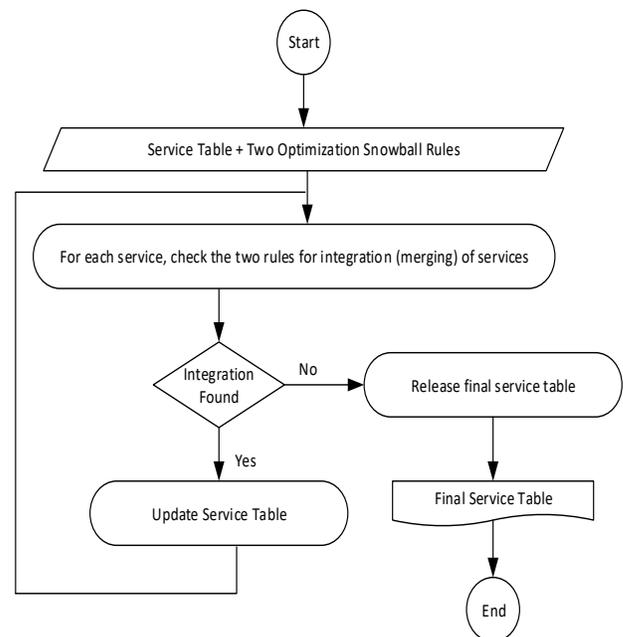


Fig. 3. Step 3: The Optimization Process.

IV. EVALUATION

To proof the significant enhancement of the proposed framework we focused on two main factors that affect SOA applications. The two factors are performance and efficiency. These two factors provide a clear measurement value that reflects how optimal a service granularity is. Performance is measured by calculating the response time between a service invocation and its reply. Efficiency is measured by calculating the total size of message used during the process of service request and reply. An insurance company was selected to evaluate the proposed work because insurance applications combine different functionalities from different business domains.

A. Case Description

The insurance company has many valuable software assets that support many of its running business processes. The company would like to offer the existing software functionalities and any newly created ones as web services. The focus is on the SOA-application that support car insurance claim BP, whereby a client request insurance reimbursement for his/her car. To prove the significant enhancement of the proposed framework the system functionalities are built using two approaches. In the first approach the system is designed using traditional SOA approach. In the Second approach, the required functionalities are built using Snowball framework. For the two approaches the response time and the total size of data transferred (message size) are calculated starting from receiving a user request till the claim process is completed.

B. Applying Traditional Method and Snowball Framework

For a traditional approach, each one of the listed in Table I one and representing different business tasks is mapped to an individual web service. Then a service portfolio is constructed without regarding the service granularity.

For snowball, the three-step processes mentioned above were executed as follows:

Step 1: The first step takes the car insurance claim BP as input and breaks it down to elementary tasks, which results in a task table, as shown in Table I.

Step 2: The second step takes the task table as input and maps each of the tasks to an elementary web service with single operation according to the first set of rules, as shown in Fig. 2.

Table II shows the resulting services. Each service describes the mapped task, the input and output parameters, the database to which it is connected, and its dependency to other services.

Step 3: The third step takes the service table as input and performs optimization according to the second set of rules, as shown in Fig. 3. The optimization is a recursive, where each iteration updates the service table according to the new integrated services. The process ends when there are no more services to be integrated. Every service in the service table is tested whether it can be integrated with another service according to the five factors defined in section 3.B.

Applying such factors to the service table presented in Table II results in the following integrations, as shown in Table III.

The three services S2, S3, and S4 require the same input: client ID, client name, and insurance document ID. Moreover, the three services are connected to the same databases: client database and the insurance document database. The three services also have control dependency as they all should be executed sequentially before S5 is invoked. Accordingly, the three services S2, S3, and S4 are aggregated into one service, named Sa. The three services S6, S7, and S8 can also be aggregated together, as they have client ID as an input parameter. The three services are also connected to the same database that is cars database. S6, S7, and S8 should be executed as prerequisite condition for S9 and S11. Accordingly, the services S6, S7, and S8 are aggregated into one service, named Sb.

Table III shows the final services produced by Snowball. These are S1, S5, Sa, Sb, S9, S10, S11, and S12. For each integrated service, Table III describes the service functionality, the input and output parameters, the database to which it is connected, and its dependency to other services.

C. Experimental Results

The insurance application is built as a web services-based application by two different approaches: SOMA and Snowball.

- Traditional SOA application build up using 12 separated services.
- Snowball designed application that is built up using only 8 services after integrating the dependent services.

1) *The experiment:* The two applications were built using C# in .Net environment. The services were implemented on IBM server with processor Xeon E5, whereas, the service invocations were applied from a desktop with CPU core i7 and 16 M Byte memory. SOAP-UI was used as a testing tool to calculate the message size and the response time. The experiment was repeated ten (10) times and the average value was calculated. The response time and message size were calculated.

TABLE I. THE ELEMENTARY TASK OF THE BP

- Receive a claim
- Check insurer payment
- Check whether the claim is in the insured period
- Check whether the insurer have many claims (manipulator)
- Take a decision for the claim
- Get car year
- Get car making company
- Get car model
- Get car price
- Calculate estimated cost
- Get new car cost
- Get a decision for payment
- Payment

TABLE II. RESULTING SERVICES SHOWING THE ELEMENTARY TASK OF THE BP

Service	Functionality	Input parameters	Output parameters	Connected Database	Dependent services
S1	Receive a claim	- Request from client			
S2	Check insurer payment	- -Client ID - -Client name - -Insurance document ID	- Boolean value	Client DB Insurance Document DB	S3, S4
S3	Check whether the claim is in the insured period	- -Client ID - -Client name - -Insurance document ID.	- Boolean value	Client DB Insurance Document DB	S2, S4
S4	Check whether the insurer has many claims (manipulator)	- -Client ID - -Client name - -Insurance document ID.	- Boolean value	Client DB Insurance Document DB	S2, S3
S5	Take a decision for the claim	- -Three Boolean values	- Boolean value		S2, S3, S4
S6	Get car year	- Car ID	- Car year	Cars DB	-
S7	Get car making company	- Car ID	- Making company	Cars DB	-
S8	Get car model	- Car ID	- Car model	Cars DB	-
S9	Get damaged component prices	- -Car year - -Car making company - -Car model	- Damaged component price	External DB	S6, S7, S8
S10	Calculate estimated maintenance cost	- -Damaged component price - -Maintenance cost	- Maintenance cost		S9
S11	Get new car cost	- -Car year - -Car making company - -Car model	- New car cost	-External DB	S6, S7, S8
S12	Get payment decision	- -Maintenance cost - -New car cost	- String		S10, S11

TABLE III. FINAL SERVICES COMPOSITION BY SNOWBALL FRAMEWORK

Service	Functionality	Input parameters	Output parameters	Connected Database	Dependent Services
S1	Receive a claim	Request from client			
Sa= integration of S2, S3, S4	Check insurer payment: check whether the claim is in the insured period and whether the insurer has many claims (manipulator)	- Client ID - Client name - Insurance document ID	- 3 Boolean values	- Client DB - Insurance Document DB	
S5	Take a decision for the claim	- 3 Boolean values	Boolean value		Sa
Sb= integration of S6, S7, S8	Get car year, car making company, and car model	- Car ID	Car year Making company Car model	- Cars DB	
S9	Get damaged component price	- Car year, - Car making company - Car model	Damaged component price	- External DB	Sa
S10	Calculate estimated maintenance cost	- Damaged component price - Maintenance cost	Maintenance cost		S9
S11	Get new car cost	- Car year - Car making company - Car model	New car cost	- External DB	Sb
S12	Get payment decision	- Maintenance cost - New car cost	String		S10, S11

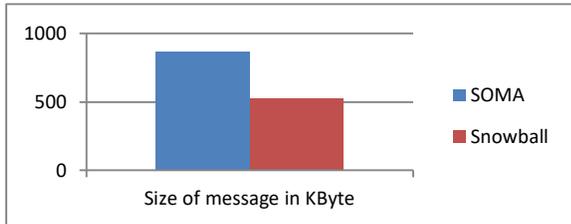
- Reply time: calculated using SOAP-UI

2) *Results:* The results listed in Table IV show a significant enhancement in the message size and respond time while maintaining the same flexibility and reusability features.

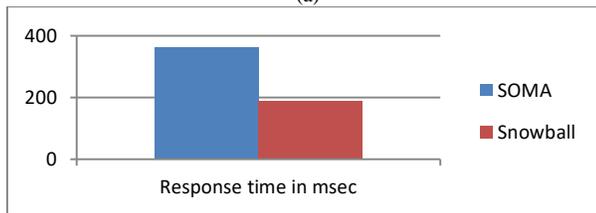
Fig. 4 shows the difference between traditional application and Snowball application across three metrics: message size (Fig. 4a), response time (Fig. 4b), and database connection (Fig. 4c).

TABLE IV. EXPERIMENTAL RESULTS

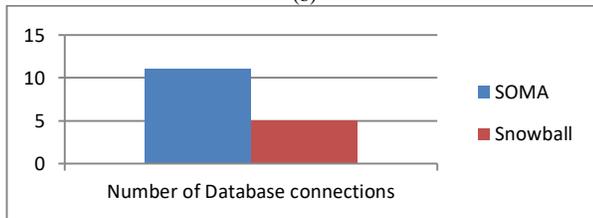
Approach	Size of transferred message	Response time	Number of connected database
Traditional SOMA	850 Kbyte	370 msec	11 databases
Snowball framework	510 Kbyte	190 msec	5 databases



(a)



(b)



(c)

Fig. 4. (a) Total Message Size in Kbyte, (b) Total Response Time in Msec, (c) Number of Connected Databases.

D. Discussion

The experimental results show a significant enhancement in two important parameters that affect the SOA applications. The first parameter is the performance that is measured by the response time of the invoked service. The average response time decreased by 48% from 362 in case of traditional applications, to be 188 msec using Snowball framework. The second parameter is the application efficiency that is measured by the size of transferred data. The transferred data significantly decreased by 39 % to be 530 KB in snowball application rather than 869 KB in case of traditional application. Such decrement in data size provides better network utilization specially when using wireless connections with narrow band width. Another significant enhancement was decreasing the number of connections between web services and database to only 5 databases rather than 11 in traditional applications.

Such results proof the significant enhancement of SOA applications while using snowball approach. Using snowball frame work defines an optimal service granularity that significantly decreases the application response time and its total transferred l data size.

It is worth mentioning that the experiment is meant to evaluate three criteria in a specific testing configuration: response time, size of the message, and number of DB connections. It also considers only one running BP to clarify our idea without adding more complexity in designing and implementing more BP.

Adding other metrics that measure a quality SOA-based application may result in a tradeoff.

V. CONCLUSION

The granularity of individual services that compose web service-based SOA is an important issue. Service granularity has significant impacts on the quality of the services regarding performance and efficiency. This work has tackled the issue of how to find an optimal service granularity. The Snowball framework was proposed to adjust web service granularity to maintain flexibility, performance, and efficiency of SOA systems. The framework is made up of two sets of rules and a three-step process.

The proposed framework was demonstrated and evaluated through the development of a web-services-based SOA application that supports the car insurance claim BP of an insurance company. The application was developed by traditional SOMA approach and Snowball framework. The experimental results show significant enhancement of Snowball over SOMA, in terms of response time, message size and DB connections. The snowball is limited to the optimization of the granularity from the perspective of performance of the services and the applications composed out of them.

The framework has practical and theoretical impacts. The developers of SOA-based applications can use it to optimize the granularity of their services to enforce their reuse and consequently the time to market. From, a theoretical perspective, the proposed work opens issues related to the optimization of the service granularly with respect to other quality criteria.

The future work will discuss the security enhancement offered by Snowball approach. Also our future work will analyze the problems associated with web service run time failure while using Snowball approach.

REFERENCES

- [1] Mohsen, A. and Naeem, K., "A review and future directions of SOA-based software architecture modeling approaches for System of Systems," In Service Oriented Computing and Applications, Volume 12, Issue 3-4, pp 183-200.2018.
- [2] Pulparambil, S., and Baghdadi, Y., "Service oriented architecture maturity models: A systematic literature review," Computer Standards & Interfaces, 61, 65-76. (2019).
- [3] Papazoglou, M. P. and Georgakopoulos, A. D., "Service-Oriented Computing," In Communications of the ACM, vol. 46 (10), pp. 24-28. 2003.
- [4] Erol, O., Mansouri, M., and Sauser, B., "A framework for enterpriseresilience using service oriented architecture approach," In: 20093rd annual IEEE systems conference. IEEE, pp 127-132.2009.
- [5] Baghdadi, Y., "Modelling business process with services: towards agile enterprises," International Journal of Business Information Systems, 15(4), 410-433. 2014.

- [6] Baghdadi, Y. and Al-Bulushi, W. "A guidance process to modernize legacy applications for SOA," SOCA 9, 41–58 . <https://doi.org/10.1007/s11761-013-0137-3>.2015.
- [7] Baccar, S. Baccar, S., Rouached, M., Verborgh, "Declarative Web service composition using proofs," In Service Oriented Computing and Applications Volume 12, Issue 3–4, pp 371–389. 2018.
- [8] Immonen, A. and Pakkala, D., "A survey of methods and approaches for reliable dynamic service compositions," In Service Oriented Computing and Applications, Volume 8, Issue 2, pp 129–158.2014.
- [9] Ayed Alwadain, Erwin Fieft, Axel Korthaus, Michael Rosemann, "Empirical insights into the development of a service-oriented enterprise architecture," Data & Knowledge Engineering, Volume 105, Pages 39–52, ISSN 0169-023X.2016.
- [10] Ding, Z. and Zhou, Z., "Race Test: harmful data race detection based on testing technology in WS-BPEL," In Service Oriented Computing and Applications 13:141–154 doi.10.1007/s11761-019-00261-1.2019.
- [11] Silic, M., Delac, G., and Srblic, S., "Prediction of atomic web service reliability based on k-means clustering," In proceedings of the 9th Joint Meeting on Foundations of Software Engineering, pages 70–80, Saint Petersburg, Russia — August 18 - 26, ISBN: 978-1-4503-2237-9 doi>10.1145/24914112491424.2013.
- [12] Baghdadi, Y., "SOA Maturity Models: Guidance to Realize SOA," International Journal of Computer and Communication Engineering 3 : 372-378, DOI:10.7763/IJCE. 2014.V3.352.2014.
- [13] A. Arsanjani, L. Zhang, M. Ellis, A. Allam and K. Channabasavaiah, "S3: A Service-Oriented Reference Architecture," in IT Professional, vol. 9, no. 3, pp. 10-17, doi: 10.1109/MITP.2007.53. May-June 2007.
- [14] Osshiro M. et al. Márcio Osshiro, Elisa Y. Nakagawa, Débora M. B. Paiva, Geraldo Landre, Edilson Palma, Maria Istela Cagnin, "Cambuci: A Service-Oriented Reference Architecture for Software Asset Repositories," In: Latifi S. (eds) Information Technology - New Generations. Advances in Intelligent Systems and Computing, vol 558, Springer, Cham https://doi.org/10.1007/978-3-319-54978-1_74 ISBN 978-3-319-54978-1.2018.
- [15] Laradi, N., Bernard, P. and Plaisent, M., "The Organizational Impacts of a Service Oriented Architecture," In Journal of Economic Development, Management, IT, Finance and Marketing, 10(1), 88-96, March 2018.
- [16] X. Liu, Y. Ma, G. Huang, J. Zhao, H. Mei and Y. Liu, "Data-Driven Composition for Service-Oriented Situational Web Applications," in IEEE Transactions on Services Computing, vol. 8, no. 1, pp. 2-16, , doi: 10.1109/TSC.2014.2304729. Jan.-Feb. 2015.
- [17] Camacho, J. A., Chamorro, C. D. and Caicedo, N. G., " Implementation by means of Web service with Service Orientation Architecture for a System Tele-operated." In International Seminar of Biomedical Engineering (SIB) IEEE: 10.1109/SIB.2018.8467754, 16-18 May, Bogota, Colombia. 2018.
- [18] Candido, G. Colombo, A., Barata, J. and Jammes F, "Service-oriented infrastructure to support the deployment of evolvable production systems," In IEEE Trans. Ind. Informat., vol. 7, no. 4, pp. 759-767. 2011.
- [19] Razavian, M., Lago, P. A., "Systematic Literature Review on SOA Migration," In Journal of Software: Evolution and Process, 27(5), 337-372. <https://doi.org/10.1002/smr.1712>.2015.
- [20] Erl, T., "SOA Principles of Service Design (paperback)," The Book. Prentice Hall Press Upper Saddle River, NJ, USA ISBN:0134695518 9780134695518.2016.
- [21] Choi, D.L. Nazareth, H.K. Jain, "The impact of SOA implementation on IT-business alignment: a system dynamics approach," In ACM Trans. Manag. Inf. Syst. 4, 3 (<https://doi.org/10.1145/2445560.244556>). 2013.
- [22] Baryannis, G., Kritikos, K. and Plexousakis, D., "A specification-based QoS-aware design framework for service-based applications," In Service Oriented Computing and Applications, 11: 301.doi10.1007/s11761-017-0210-4.2017.
- [23] Niknejad, N., Hussin, A.R.C., and Amiri, I.S., "Introduction of Service-Oriented Architecture (SOA) Adoption," In: The Impact of Service Oriented Architecture Adoption on Organizations. Springer Briefs in Electrical and Computer Engineering. Springer, Cham.2019.
- [24] Guinard, D., Trifa, V., Karnouskos, S., Spiess, P. and Savio, "D. Interacting with the SOA-Based Internet of Things: Discovery, query selection, and on-Demand Provisioning of Web service," In ServComput IEEE Trans 3(3):223–235. 2010.
- [25] Kumar, L. Kumar, S. R. and Sureka, A. (2017) Using source code metrics to predict change-prone web service: A case-study on ebay services. In 2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE)Klagenfurt, Austria IEEE, DOI:10.1109/MALTESQUE.2017.7882009.
- [26] Gazzarata, R, Vergari, F, Cinotti, T. S. and Giacomini, M., "A standardized SOA for clinical data interchange in a cardiac tele monitoring environment," In IEEE J. Biomed. Heal. Informatics, vol. 18, no. 6, pp. 1764-1774. 2014.
- [27] H. M. Sneed, "Integrating legacy software into a service oriented architecture," Conference on Software Maintenance and Reengineering (CSMR'06), 2006, pp. 11 pp.-14, doi: 10.1109/CSMR.2006.28.

Selection of Requirement Elicitation Techniques: A Neural Network based Approach

Mohd Muqem¹, Sultan Ahmad^{2*}, Jabeen Nazeer³, Md. Faizan Farooqui⁴, Afroj Alam⁵

Department of Computer Application, Integral University, Lucknow, India^{1,4,5}

Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Alkharj, 11942, Saudi Arabia^{2,3}

Abstract—Requirement Elicitation is key activity of requirement engineering and has a strong impact on design and other phases of software development life cycle. Poor requirement engineering practices lead to project failure. A sound requirement elicitation process is the foundation for the overall quality of software product. Due to criticality and high impact of this phase on overall success and failure of projects, it is very necessary to perform the requirements elicitation activities in a perfect and specific manner. The most difficult and demanding jobs during Requirement Elicitation phase is to select appropriate and specific technique from a wide array of techniques and tools. In this paper, a new approach is proposed using an artificial neural network for selection of requirement elicitation technique from a wide variety of tools and techniques that are available. The training of Neural Network is done by back propagation algorithm. The trained and resultant network can be used as a base for selection of requirement elicitation techniques.

Keywords—Requirement elicitation; requirement engineering; neural network; back propagation

I. INTRODUCTION

The process of Requirement Engineering (RE) starts with requirement gathering i.e.; requirements elicitation[1][2]. Requirement elicitation provides a base for a RE, has a strong impact on the software project success and quality [3]. The first stage in requirement engineering is requirement elicitation that uses various techniques to elicit requirements related to software project from various stakeholders. According to [4] in requirement elicitation process, the meaningful information is collected in the form of requirement from various stakeholder and users. Requirement Elicitation is the collection of requirements; the other terms used in this context are capturing of facts, knowledge acquisition, determination of essentials, gathering of information. Sometimes following terms are used such as identification, invention, development, knowledge discovery, and fact-finding. The other terms that have been used are insufficient to represent real meaning, facts, derived knowledge. But now it is well understood and minutely documented and universally accepted that requirements are captured or collected, instead the requirement are elicited [5]. Another definition for requirement elicitation is that it is the process of identification of key software requirements that are elicited from various elicitation techniques such as formal interviews, brain storming, model workshops, workflow analysis and other techniques [6]. “Requirement elicitation refers to as the

process as trawling for requirements”. During requirement elicitation, requirements are elicited in consultation with different stakeholders. The stakeholders, as they are distinct individuals, they have their specific means to recognize and store the knowledge. They express their knowledge about the problem domain in their unique way; a single method as suggested by other researchers will not be sufficient enough to elicit requirements in requirement elicitation phase from different sets of stakeholders [7]. However, it is a well-known fact that if requirements are not properly elicited then it will lead to failure of software product. To improve the success rate of software projects requirement elicitation process should be followed minutely and effectively [8]. Requirement Elicitation can be considered as a prerequisite for all other software development activities. It signifies that the developed software product is of good quality and will work properly according to the expectation [3]. The success and failure of the project is dependent on the quality of requirements that are gathered. Software’s project success and failure depends upon requirement elicitation specially selection of requirement elicitation techniques [9]. It is established fact that process of software requirement elicitation has a major and decisive impact on quality of the final product [10][11]. The main goal of effective requirement engineering is to develop a software product in such a way that fulfills all the customer needs budget, cost and schedule. Many surveys and research have been analyzed to investigate the major cause of software project failure and their statistics. Standish reports [12] suggest that rate of success of software project is 28% only. The major factor for such low rate of software successful software is imprecise and unclear requirements [13]. Another survey suggests that only 12.7% out of (nearly) thousand software projects were successful. According to survey main reason for the software project failure was unclear and incomplete requirements [14]. These survey and reports also suggest that for successful requirement engineering, requirement elicitation phase should be carried out in such a way that all the effective requirements are gathered which lead to the successful software projects. In [15] author suggests that 90% of large-scale software projects are failure due to poor requirement engineering specially requirement elicitation. This work was reflection of [16] according to author “poor software requirements management can lead to 71% percent of software projects failure; greater than badly used technologies, slipped deadlines and change management. The cost of software project failure is high”. Researchers also suggest that 70% of the total software errors and bugs are due

*Corresponding Author.

to poor requirements gathering and remaining 30% are due to poor and faulty design. It is well documented from various surveys and researchers that requirement elicitation phase has a major impact on software product quality [17]. The most important and critical task in requirement elicitation is requirement elicitation technique selection from a wide varieties of techniques[[18]. Hence there is need to understand requirement elicitation techniques, also the application of requirement elicitation techniques for successful requirement elicitation in requirement engineering phase of SDLC [19].

Many frameworks define requirement elicitation process, or describe specific elicitation technique which is performed during requirement elicitation. But none of framework proposed by various researchers has defined a unified model for the requirement elicitation process. That takes into consideration, the basic issues and the knowledge required for effective elicitation of requirements from various stakeholders. Requirement elicitation framework suggested by various researchers are theoretical and there is lack of mathematical knowledge[19]. An efficient framework is required for effective requirement elicitation. There are various techniques of elicitation but how to select the appropriate techniques from a set of techniques is the real challenge. The elicitor must decide which technique is best suited for specific situation from set of techniques [20]. Inappropriate selection of elicitation technique leads to gathering of faulty requirement which results in project failures. Because of variety of stakeholder or rather non homogeneity of stakeholders, process of requirement elicitation must be carried out effectively by applying the appropriate elicitation techniques. In this research paper important key factors are identified from various researchers and practitioners of software industries that directly or indirectly contribute towards selection of techniques of requirement elicitation. These factors are fed as input to neural network and the intelligent approach is applied for requirement elicitation technique selection from a set of various techniques and tools.

II. RELATED WORK

During requirement elicitation phase or requirement engineering the level of scope for the system need to be established. The details regarding the needs and requirement of key stakeholders should be investigated using variety of elicitation techniques. Requirement elicitation technique is related to eliciting of requirements of the organization within the organization environment including project goals, project rules, processes work flow, various assumptions market, constraints along with details of implementation. In [21] researches suggest that it is important to elicit different types of information during elicitation process. To have the complete knowledge of the systems including all the details of current and existing system in a specific domain, along with the current list of existing requirement engineering problems their goals including issues, ideas and risk. This is known as perspective of concept modeling. In this modeling elicitation of requirements is done for processes, data and behavior of large-scale software projects [22]. This model is also supported by SA/SD view of large-scale software project [23].

The important components of the elicitation process are priority, source and rationale of the software requirements along with the goals [24]. Each elicitation technique has some key features and also some limitations [25], Interview is most common and natural elicitation technique[25], there is no support for this elicitation process and even no specific guideline for eliciting requirements [26]. In researcher suggested that there is no elicitation techniques that provides detailed description of the problems [27]. Suggested that there are two types of requirement gathering techniques firstly the techniques that are less expensive as well as information provided by them is less[28][29]. Secondly the techniques those give favorable result but are expensive. The elicitor may be well versed with some elicitation techniques but no one elicitation technique is able to elicit all the requirements entirely[30]. Hence more than one requirement elicitation technique is used to capture the requirements for software systems. Because of wide range of elicitation techniques, it is therefore possible to use alternatively available requirement elicitation techniques in different scenarios with enriched flexibility, providing more options to the elicitor.

The elicitation techniques are informal in nature that deeply involves human interaction. Group work is one technique that is effective as it involves groups and is able to deal with complex scenarios such as elicitation process. This is far better than individuals because each individual is unique and possesses wide range ability. Group work techniques are better in gathering of requirement as they involve stakeholder and customers. It fosters discussion, generation of ideas and fact-finding solution. The major benefit of Requirements Workshops is that they glue well techniques of elicitation. The group techniques are important in eliciting requirements because software engineering is a group activity [31][32]. In [33] suggests that one to one interview is the most suitable method for requirement gathering. In [34] researchers suggest that workshop is most suitable technique for small scale software projects whereas [35] suggest that DSDM is suitable for large scale software projects. In [36] researchers provide guidelines for effective requirement elicitation technique selection.

The works done in [37][38][39][40] suggests novel approaches for elicitation of requirement. The above researchers proposed many process models for elicitation of requirement over the period of time [40][41][42]. Mentioned approaches provide us flexibility and take into consideration the individual software projects. The approaches do not provide guidelines as number of tasks are performed during elicitation phase. Many problems are faced by elicitor as there are many techniques available to make the software development task simple. The elicitation is carried out over different sessions to capture every detail in parallel with Software Development Life Cycle process. The author in [43] suggests that elicitation involves the understanding of both problem and business domain of the large organization and also to find out how the system works along the understanding of the application domain of the current working software project. In [44] recommends that object setting, knowledge acquisition and requirement gathering from stake holder are requirement for a standard elicitation process.

In [45] researchers suggested a step wise approach using the ISO 9126 quality characteristics as a guideline for the requirement elicitor to capture social, organizational and human factors that can improve overall quality of software product. In [46] researchers suggested the analytic network process, based on multi-criteria decision making for requirement elicitation technique selection. In [47] researcher analyzed for IT professionals the cognitive structure based on various factors that directly affect the use of the ladderings as requirement elicitation techniques selection. In [48] researcher suggests that interview is best technique for requirement elicitation technique selection. The results of elicitation are analysed using children education application as case study. The outcome of this research is interview along with prototyping method of requirement elicitation technique selection are the most suitable methods. Although there are various framework and guidelines for requirement elicitation technique selection till date but most of them are theoretical in nature and lacks mathematical concepts. We required a more effective and intelligent approach that provide researchers with detailed guidelines for carrying out elicitation technique selection process empirically.

III. PROBLEM DEFINITION AND SCOPE

The software projects failure is main issue of the software development industry. Researchers had tried to investigate the cause of failure in software projects failure statistically. Empirical research in organizations has demonstrated that requirement engineering process which is well defined has a positive impact on the quality, cost of software developed as compared to projects that do not follow well defined process models. Chaos Report suggests the problems of delay in software development, over-budget software and failed software are major concerns for software industry. To develop high quality software is also one of the concerns. Some researcher suggests that the poor requirement elicitation led to 50% of the total failure of software project. Significant efforts have been made in requirement elicitation research; there exist gap between theory and actual practice. In order to improve success rate in software developed an efficient and effective approach is needed for requirement elicitation. In Elicitation process there are number techniques available still it is difficult for requirement elicitor to decide which technique is most suited for the task [13][14]. It is due to inability to understand the available techniques. This inability to understand the technique leads to improper selection of requirement elicitation technique which leads to project failure. So, it is imperative for requirement elicitor to know about the existing requirement elicitation techniques [2][8]. Although elicitors have a number of elicitation techniques at their disposal still flexible guidelines for effective approach of requirement selection are needed, which are beneficial in selection of appropriate elicitation techniques [6].

IV. ATTRIBUTES FOR THE REQUIREMENT ELICITATION TECHNIQUE SELECTION

Selection of Elicitation techniques are depended based on various attributes. Attributes plays an important role in

selection of requirement elicitation techniques. These key attributes are identified after intensive literature survey and consultation of professionals working in different software houses. There are 9 different attributes which are identified for selection of elicitation techniques. These attributes are known as evaluating factors for selection of requirement elicitation technique. The attributes are shown in Table I.

A. Software Project Scaling

Software Project Scaling can happen in Man-hours, code-size, number of interfaces, number of requirements and cost. Scaling happens with headcount, as large numbers of developers are required to develop large software hence more developers should produce more code. The size of products is proportional to size of the code. Many products can be developed simply by rearranging the code. Or in other words several products can be developed from the single product. On the other hand, if a single-person develops the software then the cost of mistakes is usually lesser. Software developers usually take shortcuts to achieve their goals. In small software projects developer and manager is a same person. In large software projects, the roles are distributed among many persons hence there is the need for coordination. So from the above discussion it is justified that software can be classified into small scale software and large scale software. The classification is based on the following Attributes described in the Table II.

TABLE I. EVALUATING FACTORS FOR SELECTION OF REQUIREMENT ELICITATION TECHNIQUES

Attributes	Values
Domain	Understanding the Domain
	Acquiring Domain Knowledge
	Domain Characteristic identification
Stakeholder	Stakeholder Identification
	Stakeholders Classification
Elicitation	Identification of Requirement Sources
	Tool Selection
Others	Budget Constraints
	Quality concerns
	Project Status

TABLE II. PROJECT SCALING ATTRIBUTES

Attribute	Small Scale Project	Large Scale Project
Man-hours	Small	Large
Interfaces	Less	More
Requirements	Small	Large
Features	Simple	Complex
Cost	Low	High
Quality	Minimum QA	Larger QA
Deadlines	Flexible	Stringent

V. SUITABILITY ANALYSIS OF REQUIREMENT ELICITATION TECHNIQUES

There are varieties of Elicitation Techniques but the challenge is to select the right one from a wide variety of requirement elicitation techniques. To select the evaluating factors for selection of requirement elicitation techniques, literature was surveyed intensively and professionals working in different software houses were consulted.

After literature survey and discussion with professionals, 9 evaluating factors are identified which play important role while requirement elicitation technique selection. The impact of evaluating factor for selection of requirement elicitation techniques are studied and documented for small scale & large-scale projects. The selection of requirement elicitation techniques in small scale and large-scale projects based on the evaluating factors are shown below in Table III and Table IV.

TABLE III. SELECTION OF ELICITATION TECHNIQUES IN SMALL SCALE PROJECTS

Requirement Elicitation Technique									
Evaluating Factors	Interface prototyping	Observation	Brainstorming	Interview	Group Meetings	Workshop	Questionnaires	Expert Interviews	Scenarios
Understanding the Domain	x			x					x
Acquiring Domain Knowledge		x		x			x		
Domain Characteristic identification		x	x		x			x	
Stakeholder Identification				x		x			x
Stakeholders Classification			x	x	x			x	
Identification of Requirement Sources		x	x	x		x			
Tool Selection	x			x			x		x
Budget Constraints		x	x		x	x			
Project Status			x	x			x		x

TABLE IV. SELECTION OF ELICITATION TECHNIQUES IN LARGE SCALE PROJECTS

Requirement Elicitation Technique									
Evaluating Factors	Interface prototyping	Observation	Brainstorming	Interview	Group Meetings	Workshop	Questionnaires	Expert Interviews	Scenarios
Understanding the Domain	x	x		x		x			
Acquiring Domain Knowledge			x	x		x	x		x
Domain Characteristic identification		x		x	x	x			
Stakeholder Identification				x	x	x	x		x
Stakeholders Classification			x	x	x		x		
Identification of Requirement Sources	x		x	x		x			x
Tool Selection	x			x					
Budget Constraints	x							x	x
Project Status			x				x	x	x

VI. SELECTION OF REQUIREMENT ELICITATION TECHNIQUES: A NEURAL NETWORK BASED APPROACH

Neural Networks possess the computing capability of the brain of human. The neural network consists of computational units that are functionally similar to neurons that are there in the human brain. A neural network is composed of several layers of neurons. Every neuron performs calculative work that will contribute to the learning of the neural network. Neurons consist of activation function, synaptic weights and an adder. If x_j is the input to neuron named i then w_{ij} is the synaptic weight that is related with neuron and input signal. In a multi-layer feed forward neural network there is a layer that is used for input for non-computational units. There are several layers of computational units that are hidden. Lastly there is output layer that consists of computational units. For the training of multi-layer feed forward network, Back propagation is used. This algorithm forwards the input signal and backward the error signal through the network. Multi-layer feed forward neural network solves complex problems and back propagation method is used to train the network.

In requirement elicitation the requirements are elicited from various sources using various elicitation techniques. There are wide varieties of elicitation techniques the major challenge is to select the appropriate elicitation technique which provides support to the elicitor and also understanding their needs and requirements [7]. Many of the elicitor think that only one of the elicitation technique is used in all types of projects and applicable to all situations, but one elicitation technique is not sufficient for all types of projects and situations [1][2][3][4][5]. In requirement elicitation elicitor selects a particular elicitation technique from a set of techniques for one reason or the other. (1) The elicitor is well versed with only one specific technique. (2) In all the situations the elicitor's favours that technique. (3) The elicitor uses a particular methodology, and the methodology is supported by a particular technique. (4) The elicitor understands spontaneously that the elicitation technique is efficient for current circumstance. From the above discussion it is evident that the fourth reason is mostly supported by the elicitor. Here neural network approach is applied for selection of requirement elicitation technique and implemented using MATLAB NNtool [11] as shown in Fig. 1.

A. Proposed Algorithm: Neural network-based model for Requirement Elicitation Technique:

RequirementElicitation()

```
{
Input: (elic1, elic2, . . . ,elicn) // For each project r1,r2...rn are new set of
requirements gathered using requirement elicitation technique.
```

```
elic (Ri, Prs, Ti) //Each step in Requirement elicitation process is expressed
as a combination of elicited requirements, situation of project and
elicitation technique.
```

```
Ri= Ri+1;
```

```
Prs=Prsi+1;
```

```
// For each step i of the elicitation, elicitation technique ti when situation Si
exists and Ri captures the current state of requirement. The outcome of this
step is a new requirement with new project situation.
```

```
Ti ∈ Et //The selected elicitation technique is Ti which one is selected from
a group of requirement elicitation techniques
```

```
Ei. // Pr is Project whose requirements are elicited by applying elicitation
technique.
```

```
For the selected project Pr requirement elicitation techniques are selected
on the basis of Project attributes A represented by function.
```

```
£(Ri ,Prsi,Efv(Et)) : C ∈ Et Where Et ∈ T // Efv represents evaluating factors
on the basis of these factors elicitation techniques are selected.
```

```
Ri = Et (Pri) ∪ Req where Reqi = {Req1, Req2 ....Reqn}
```

NNReqElic() // A Neural network based approach is for selection of Elicitation Technique

- ```
{
1. v: Assigned as Input vector //
2. t: Assigned as Target values
3. Feed Forward Neural Network is created
4. Neural network is trained using Back-propagation
4. Result of training is obtained
6. Neural network is simulated
7. Results are analyzed
}
```

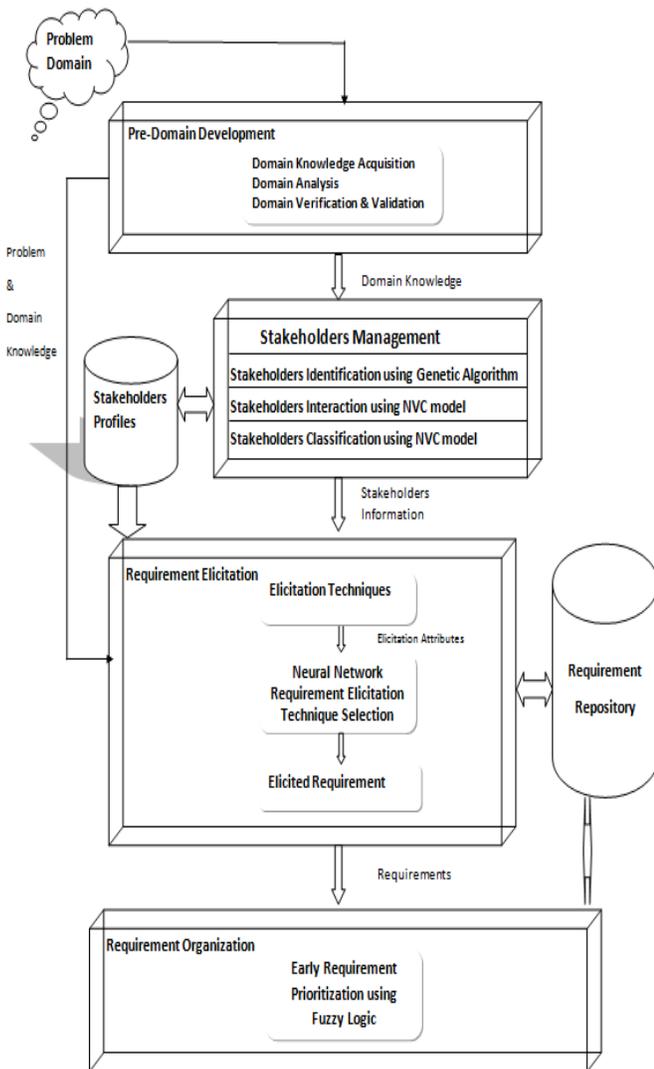


Fig. 1. Proposed Framework for Requirement Elicitation.

**B. Description of Algorithm**

Method followed for selection of Requirement Elicitation Technique.

1) Table III and Table IV of requirement elicitation techniques along with their impact are analyzed for both small scale projects & large-scale projects.

2) Matrix A is obtained from Table III and matrix B is obtained from Table IV.

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

3) Using MATLAB for Requirement Elicitation Technique selection.

- a) Neural network is created.
- b) Input vectors (v) values are Assigned.
- c) Assign target values (t).
- d) Create neural network.
  - i) Name: Requirement Elicitation.
  - ii) Network properties are used as per neural network description.
- e) Neural Network for Selection of Requirement Elicitation Technique.
- f) Training of neural network.
- g) The result of training.
- h) Simulating the neural network.
- i) Result of neural network-based approach.

The proposed neural network-based approach for selection of Requirement Elicitation Technique is implemented using MATLAB tool. The algorithm is implemented using NNTool box. The NNTool box is used for creating neural network. The working of MATLAB NNTool for selection of requirement elicitation technique is shown in Fig. 2, 3(a) and 3(b).

Input vector(v) are assigned following values  $V = \{1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1; 1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0\}$ .

Target values (t) are assigned as following values  $T = \{1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1\}$ .

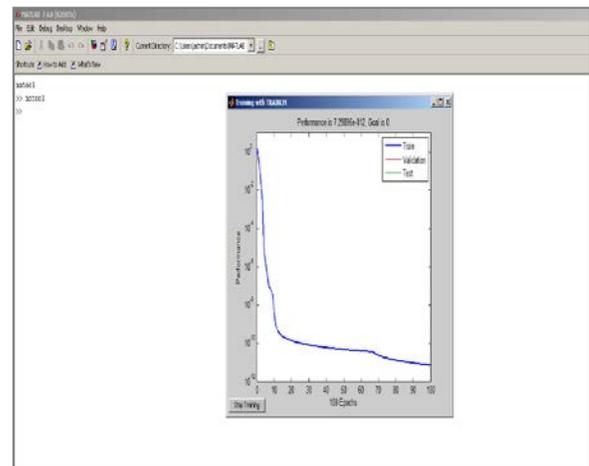
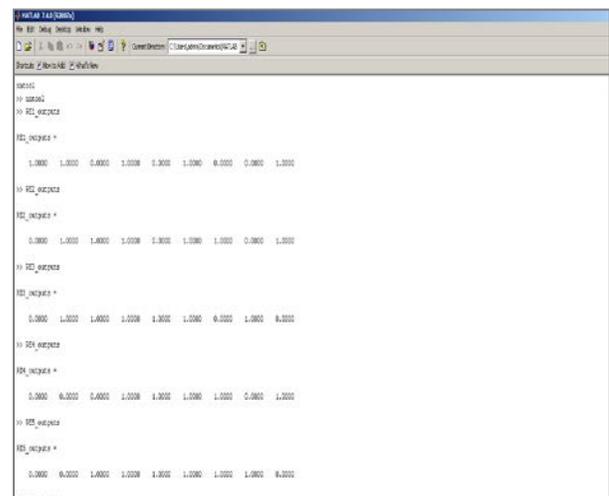


Fig. 2. Training Result.



(a)



(b)

Fig. 3. (a). Output of Neural Network (Part-1). (b). Output of Neural Network (Part-2).

The neural network (Requirement Elicitation) is created. Feed forward Back propagation is the network type. The input range is [0,1;0,1] and the Training Function TRAINLM. LEARNM is used as Adaptation learning function. MSE is used as performance function total no of layers is 2. Finally, the requirement elicitation neural network is trained and training results are generated. Requirement elicitation neural network is simulated.

VII. ANALYSIS OF RESULT

The Result is analyzed as ‘1’ in the output vector will corresponds to the elicitation technique being used for the evaluating factor of requirement elicitation and ‘0’ in the output vector corresponds to the technique being discarded for the given evaluating factor of requirement elicitation as shown in Matrix R. The result of neural network for selection of requirement elicitation technique is shown in matrix R

$$R = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

From the above findings weight count (Wc) is assigned to each of the corresponding elicitation technique as shown in the Table V. A scale of 2-10 is assigned to each technique. The technique with highest weight count will be the most effective technique of requirement elicitation and lowest count is least effective.

Results shown in Table V conclude that Interview is the most effective technique for requirement elicitation and workshop is the second efficient technique for requirement elicitation. Interview is most effective technique that starts with a set of pre-defined questions. The most preferred and easy technique to gather requirements from different set of stakeholders is Interview. Interview is the elicitation technique which provides a support to understand the problem domain of the existing system. The most formalized and pragmatic technique for requirement elicitation is Workshop and brainstorming. Brainstorming and workshop session provide ideas that are well suited on voting technique and/or other criteria. The most expensive technique for requirement elicitation is Prototyping as compared to other elicitation techniques. The best technique for data collection from a mass variety of sources is Questionnaire which is also cost effective. The most recommended elicitation technique for complex and critical requirements is Scenario. This will help to remove any ambiguity in the requirements. Scenarios are also effective for the projects which have to be completed in short time window or the projects in which process improvement is needed. There are various elicitation

techniques out of these techniques the elicitor selects the techniques that fit perfectly into the software projects. In this paper an approach is proposed for requirement elicitation technique selection and used this as base for selecting the combination of elicitation techniques that is best suited for effective requirement elicitation. The study suggests that single elicited techniques are generally used for requirement elicitation process, but combination of elicitation techniques for requirement elicitation process gives better result. There are various combinations of techniques denoted by  $T_{c1}, T_{c2}, T_{c3} \dots T_{cn}$ , the combination with high score as defined in the neural network approach for selection of requirement elicitation technique which is best suited for efficient requirement elicitation. As per the findings and the scores calculated in Table VI, appropriate techniques selection is done. From the findings in most cases,  $T_{c2}$  the techniques combination with the highest score is selected according to the case study. If there are no other factors that are to be considered in the final decision making then  $T_{c2}$  will be the final choice of the requirement elicitation techniques selection. These combined requirement elicitation techniques are selected for effective requirement gathering in software projects.

If new evaluating factor Ev added in further studies of projects the condition still holds the same.

TABLE V. WEIGHT COUNT OF ELICITATION TECHNIQUES

| Techniquet          | Weight Count Wc |
|---------------------|-----------------|
| Observation         | 5               |
| Brainstorming       | 8               |
| Interview           | 10              |
| Group Meetings      | 2               |
| Workshop            | 9               |
| Questionnaires      | 6               |
| Ethnography         | 3               |
| Scenarios           | 7               |
| Interface Prototype | 4               |

TABLE VI. COMBINATIONS OF ELICITATION TECHNIQUES AND THEIR SCORES

| Technique Combination $T_{c,t}$ | Techniquet                                         | Scoret |
|---------------------------------|----------------------------------------------------|--------|
| $T_{c1}$                        | Observation, Brainstorming, Interview              | 23     |
| $T_{c2}$                        | Brainstorming, Interview, Workshop                 | 27     |
| $T_{c3}$                        | Interview, Scenario, Interface Prototype           | 21     |
| $T_{c4}$                        | Group Meetings, Workshop, Interview                | 21     |
| $T_{c5}$                        | Workshop, Ethnography, Questionnaire               | 18     |
| $T_{c6}$                        | Questionnaires, Interface Prototype, Group Meeting | 12     |
| $T_{c7}$                        | Interface Prototype, Workshop, Observation         | 18     |

A. Statistical Analysis of Proposed Approach

In addition to the theoretical validation, an experimental statistical analysis is equally important in order to make the claim acceptable. In view of this fact, a statistical validation is performed to assess the performance of the proposed framework. One way ANOVA test is implemented using SPSS tool for statistical validation of requirement elicitation technique selection approach. In one way ANOVA only one factor is considered and investigated to find the differences amongst its various sample categories having numerous possible values. In this work evaluating factors Ev are used as factors defined in the table. There are total 9 factors and their values are shown in the Tables III and IV.

SPSS tool is used for implementing ANOVA test on two set of Sample data. One set sample data for small scale project and other set of sample data for large project considering evaluating factors. Result obtained from small scale projects while considering three evaluating factor Ev for selection of requirement elicitation technique and for large scale projects considering six evaluating factors for the same. The conclusion drawn from the results shown in Fig. 4 and 5, is that in samples (small scale & large scale project) the value of  $p < .05$  (For small scale projects  $p = .009$  and for large scale projects  $p = .001$ ) and the value of F in both samples is greater than the tabular value ( $F = 8.575$  for small scale and  $F = 7.713$  for large scale) means that there is significance difference between the groups. If there is significance difference between groups in small scale and large-scale projects, then there is sufficient evidence to prove that evaluating factors defined in both types of projects for selection of requirement elicitation technique are sufficient.

| ANOVA          |                |    |             |       |      |
|----------------|----------------|----|-------------|-------|------|
| Technique      | Sum of Squares | df | Mean Square | F     | Sig. |
| Between Groups | 81.088         | 5  | 16.218      | 7.713 | .001 |
| Within Groups  | 27.333         | 13 | 2.103       |       |      |
| Total          | 108.421        | 18 |             |       |      |

Fig. 5. ANOVA Test Result (Large Scale Project).

VIII. CONCLUSION

In requirement elicitation there are varieties of elicitation techniques but the challenge for the elicitor to select the effective technique or a combination of technique which is most suited for problem domain, stakeholders, project type, organizational structure and project to be developed. In this paper an approach using neural network is proposed for selection of requirement elicitation techniques from a large array technique. If the elicitor follows the proposed approach the problem related to selection of elicitation technique is minimized and effective requirements are elicited.

ACKNOWLEDGMENT

We thank the Deanship of Scientific Research, Prince Sattam Bin Abdulaziz University, Alkharij, Saudi Arabia for help and support.

REFERENCES

- [1] J. Li et al., "Attributes-based decision making for selection of requirement elicitation techniques using the analytic network process," Math. Probl. Eng., vol. 2020, 2020.
- [2] A. Hussain and E. O. C. Mkpojiogu, "Requirements: Towards an understanding on why software projects fail," in AIP Conference Proceedings, vol. 1761, no. 1, p. 20046, 2016.
- [3] A. Saad and C. Dawson, "Requirement elicitation techniques for an improved case based lesson planning system," J. Syst. Inf. Technol., 2018.
- [4] Z. Wang, C.-H. Chen, P. Zheng, X. Li, and L. P. Khoo, "A novel data-driven graph-based requirement elicitation framework in the smart product-service system context," Adv. Eng. informatics, vol. 42, p. 100983, 2019.
- [5] D. A. Tamburri, "Design principles for the General Data Protection Regulation (GDPR): A formal concept analysis and its evaluation," Inf. Syst., vol. 91, p. 101469, 2020.
- [6] N. C. Alflen, E. P. V Prado, and A. Grotta, "A Model for Evaluating Requirements Elicitation Techniques in Software Development Projects.," in ICEIS (2), pp. 242–249, 2020.

| Descriptives |    |        |                |            |                                  |             |         |         |  |
|--------------|----|--------|----------------|------------|----------------------------------|-------------|---------|---------|--|
| Technique    | N  | Mean   | Std. Deviation | Std. Error | 95% Confidence Interval for Mean |             | Minimum | Maximum |  |
|              |    |        |                |            | Lower Bound                      | Upper Bound |         |         |  |
| 1            | 4  | 2.7500 | .85743         | .47671     | 2.2825                           | 3.2175      | 2.00    | 5.00    |  |
| 2            | 3  | 5.3333 | 1.52753        | .88182     | 4.5088                           | 6.1580      | 4.00    | 7.00    |  |
| 3            | 4  | 7.2500 | 2.15500        | 1.08250    | 6.16750                          | 8.33250     | 6.00    | 9.00    |  |
| Total        | 11 | 5.3636 | 1.74773        | .53686     | 4.29358                          | 6.43362     | 3.00    | 9.00    |  |

| ANOVA          |                |    |             |       |      |
|----------------|----------------|----|-------------|-------|------|
| Technique      | Sum of Squares | df | Mean Square | F     | Sig. |
| Between Groups | 21.129         | 2  | 10.564      | 0.075 | .009 |
| Within Groups  | 8.417          | 8  | 1.052       |       |      |
| Total          | 29.546         | 10 |             |       |      |

Fig. 4. ANOVA Test Result (Small Scale Project).

- [7] S. Sharma and S. K. Pandey, "Revisiting requirements elicitation techniques," *Int. J. Comput. Appl.*, vol. 75, no. 12, 2013.
- [8] S. Kaul, "Efficient Decoder for Optical Transport Networks Achieving Near Capacity Performance." 2019.
- [9] A. Y. Aleryani, "The Impact of the User Experience (UX) on the Quality of the Requirements Elicitation," *Int. J. Digit. Inf. Wirel. Commun.*, vol. 10, no. 1, pp. 1–10, 2020.
- [10] C. Giardino, N. Paternoster, M. Unterkalmsteiner, T. Gorschek, "Software development in startup companies: the greenfield startup model," *IEEE Trans. Softw. Eng.*, vol. 42, no. 6, pp. 585–604, 2015.
- [11] M. Burinskienė, V. Bielinškas, A. Podviezko, "Evaluating the significance of criteria contributing to decision-making on brownfield land redevelopment strategies in urban areas," *Sustainability*, vol. 9, no. 5, p. 759, 2017.
- [12] B. Davey and K. R. Parker, "Requirements elicitation problems: a literature analysis," *Issues Informing Sci. Inf. Technol.*, vol. 12, pp. 71–82, 2015.
- [13] M. Perkusich, G. Soares, H. Almeida, and A. Perkusich, "A procedure to detect problems of processes in software development projects using Bayesian networks," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 437–450, 2015.
- [14] D. M. Fernández and S. Wagner, "Naming the pain in requirements engineering: A design for a global family of surveys and first results from Germany," *Inf. Softw. Technol.*, vol. 57, pp. 616–643, 2015.
- [15] K. Kaur, P. Singh, and P. Kaur, "A Review of Artificial Intelligence Techniques for Requirement Engineering," *Comput. Methods Data Eng.*, pp. 259–278, 2021.
- [16] M. Kuutila, M. Mäntylä, U. Farooq, and M. Claes, "Time pressure in software engineering: A systematic review," *Inf. Softw. Technol.*, vol. 121, p. 106257, 2020.
- [17] J. Elijah, A. Mishra, M. C. Udo, A. Abdulganiyu, and A. A. Musa, "Survey on Requirement Elicitation Techniques: It's Effect on Software Engineering," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 5, pp. 9201–9215, 2017.
- [18] Z. Wang, C.-H. Chen, P. Zheng, X. Li, and L. P. Khoo, "A graph-based context-aware requirement elicitation approach in smart product-service systems," *Int. J. Prod. Res.*, vol. 59, no. 2, pp. 635–651, 2021.
- [19] S. N. Bhatti, M. Usman, and A. A. Jaji, "Validation to the requirement elicitation framework via metrics," *ACM SIGSOFT Softw. Eng. Notes*, vol. 40, no. 5, pp. 1–7, 2015.
- [20] N. R. Darwish, A. A. Mohamed, and A. S. Abdelghany, "A hybrid machine learning model for selecting suitable requirements elicitation techniques," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 6, pp. 1–12, 2016.
- [21] S. Elsayah, J. H. A. Guillaume, T. Filatova, J. Rook, and A. J. Jakeman, "A methodology for eliciting, representing, and analysing stakeholder knowledge for decision making on complex socio-ecological systems: from cognitive maps to agent-based models," *J. Environ. Manage.*, vol. 151, pp. 500–516, 2015.
- [22] N. Bin Ali, K. Petersen, and K. Schneider, "FLOW-assisted value stream mapping in the early phases of large-scale software development," *J. Syst. Softw.*, vol. 111, pp. 213–227, 2016.
- [23] J. J. Molwus, B. Erdogan, and S. Ogunlana, "Using structural equation modelling (SEM) to understand the relationships among critical success factors (CSFs) for stakeholder management in construction," *Eng. Constr. Archit. Manag.*, 2017.
- [24] N. Tripathi et al., "An anatomy of requirements engineering in software startups using multi-vocal literature and case survey," *J. Syst. Softw.*, vol. 146, pp. 130–151, 2018.
- [25] M. Yousuf and M. Asger, "Comparison of various requirements elicitation techniques," *Int. J. Comput. Appl.*, vol. 116, no. 4, 2015.
- [26] M. Younas, D. N. A. Jawawi, "Non-functional requirements elicitation guideline for agile methods," 2017.
- [27] M. Haleem and M. R. Beg, "Impact analysis of requirement metrics in software development environment," *ICECCT*, pp. 1–6, 2015.
- [28] S. Yang, Y. Tang, and Y. F. Zhao, "A new part consolidation method to embrace the design freedom of additive manufacturing," *J. Manuf. Process.*, vol. 20, pp. 444–449, 2015.
- [29] M. Haleem, M. F. Farooqui, and M. Faisal, "Cognitive impact validation of requirement uncertainty in software project development," *Int. J. Cogn. Comput. Eng.*, vol. 2, pp. 1–11, 2021.
- [30] Y. T. Yip, "Design research on anecdote-based knowledge elicitation for organization development," 2016.
- [31] M. Muqem and M. R. Beg, "A Fuzzy Based Approach for Early Requirement Prioritization," *Int. J. Comput. Technol.*, vol. 15, no. 2, pp. 6480–6490, 2015.
- [32] C. Palomares, C. Quer, and X. Franch, "Requirements reuse and requirement patterns: a state of the practice survey," *Empir. Softw. Eng.*, vol. 22, no. 6, pp. 2719–2762, 2017.
- [33] M. Muqem and M. R. Beg, "Validation of requirement elicitation framework using finite state machine," *ICCICCT*, pp. 1210–1216, 2014.
- [34] T. Dingsøyr, N. B. Moe, T. E. Fægri, and E. A. Seim, "Exploring software development at the very large-scale: a revelatory case study and research agenda for agile method adaptation," *Empir. Softw. Eng.*, vol. 23, no. 1, pp. 490–520, 2018.
- [35] N. Mahendra and M. Muqem, "An Approach for the Inception of Security Testing in the Early Stages of Software Development," in *CCTES*, 2018, pp. 304–307, 2018.
- [36] S. Ouhbi, A. Idrı, J. L. Fernández-Alemán, and A. Toval, "Requirements engineering education: a systematic mapping study," *Requir. Eng.*, vol. 20, no. 2, pp. 119–138, 2015.
- [37] N. Mahendra and M. Muqem, "Framework for Testing the Security of Application Software at Design Phase."
- [38] E.-M. Schön, J. Thomaschewski, and M. J. Escalona, "Agile Requirements Engineering: A systematic literature review," *Comput. Stand. Interfaces*, vol. 49, pp. 79–91, 2017.
- [39] N. Mahendra and M. Muqem, "SMBC: A Security Grading Tool for Accessing the Security at Design Phase of Software Development," *Int. J. Appl. Eng. Res.*, vol. 14, no. 8, pp. 2064–2073, 2019.
- [40] N. M. Mohammed, M. Niazi, "Exploring software security approaches in software development lifecycle: A systematic mapping study," *Comput. Stand. Interfaces*, vol. 50, pp. 107–115, 2017.
- [41] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Inf. Softw. Technol.*, vol. 64, pp. 1–18, 2015.
- [42] A. X. Ali, M. R. Morris, and J. O. Wobbrock, "Crowdsourcing similarity judgments for agreement analysis in end-user elicitation studies," in *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*, 2018, pp. 177–188.
- [43] A. M. Aranda, O. Dieste, and N. Juristo, "Effect of domain knowledge on elicitation effectiveness: an internally replicated controlled experiment," *IEEE Trans. Softw. Eng.*, vol. 42, no. 5, pp. 427–451, 2015.
- [44] M. Sadiq, "A fuzzy set-based approach for the prioritization of stakeholders on the basis of the importance of software requirements," *IETE J. Res.*, vol. 63, no. 5, pp. 616–629, 2017.
- [45] S. L. Buitrón, F. J. Pino, "A model for enhancing tacit knowledge flow in non-functional requirements elicitation," *CONISOFT*, pp. 25–33, 2017.
- [46] Z. Latif, W. Lei, S. Latif, Z. H. Pathan, R. Ullah, and Z. Jianqiu, "Big data challenges: Prioritizing by decision-making process using Analytic Network Process technique," *Multimed. Tools Appl.*, vol. 78, no. 19, pp. 27127–27153, 2019.
- [47] H. M. Elhassan Ibrahim Dafallaa, N. Ahmad, M. B. Rehman, I. Ahmad, and R. Khan, "Automating Elicitation Technique Selection using Machine Learning," *Fog, Edge, Pervasive Comput. Intell. IoT Driven Appl.*, pp. 47–66, 2020.
- [48] A. Souza, B. Ferreira, N. Valentim, L. Correa, S. Marczak, and T. Conte, "Supporting the teaching of design thinking techniques for requirements elicitation through a recommendation tool," *IET Softw.*, vol. 14, no. 6, pp. 693–701, 2020.

# The Performance of Personality-based Recommender System for Fashion with Demographic Data-based Personality Prediction

Iman Paryudi<sup>1</sup>, Ahmad Ashari<sup>2</sup>, Khabib Mustofa<sup>3</sup>

Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia<sup>1</sup>

Department of Informatics, Universitas Pancasila, Jakarta, Indonesia<sup>1</sup>

Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia<sup>2</sup>

Department of Computer Science and Electronics, Universitas Gadjah Mada, Yogyakarta, Indonesia<sup>3</sup>

**Abstract**—Currently, the common method to predict personality implicitly (Implicit Personality Elicitation) is Personality Elicitation from Text (PET). PET predicts personality implicitly based on statuses written on social media. The weakness of this method when applied to a recommender system is the requirement to have minimal one social media account. A user without such qualification cannot use such system. To overcome this shortcoming, a new method to predict personality implicitly based on demographic data is proposed. This proposal is based on findings by previous researchers stating that there is a correlation between demographic data and personality trait. To predict personality based on demographic data, a personality model (rule) is needed. This model correlates demographic data and personality. To apply this model to a recommender system, another model is needed, that is preference model which connects personality and preference. These two models are then applied to a personality-based recommender system for fashion. From performance evaluation, the precision of and user satisfaction to the recommendation is 60.19% and 87.50%, respectively. When compared to precision and user satisfaction of PET-based recommender system (which are 82% and 79%, respectively), the precision of demographic data-based recommender system is lower whereas the satisfaction is higher.

**Keywords**—Implicit personality elicitation; demographic data; personality-based recommender system; personality trait

## I. INTRODUCTION

The first method to be used in a recommender system was content-based filtering which recommend items based on similarity between keywords on item description and on user's profile [1][2]. However, as it turned out, a content-based filtering has several weaknesses, one of which is its inability to distinguish the quality of items. This is because a good quality item will have the same keyword as a bad quality item [1].

Because of the glaring weakness in content-based filtering method, a new method, collaborative filtering, is used. The inability of content-based method to differentiate between different item qualities is solved by collaborative filtering by asking users to rate all the consumed items. This rating data is then used to calculate the rating of all new items [3], then this method will select top N items with highest ratings and then recommend these items along with the estimated ratings [4]. In practice, a rating-based collaborative filtering also has several

weaknesses, one of which is cold start problem or the new user problem [5]. This issue occurs when a recommender system is unable to provide a new user with accurate recommendations, because a new user does not have a record of what items has been consumed and the rating (the user profile is still empty).

To deal with the cold start problem, a user profile must be made available as soon as a new user becomes a member of a recommender system. The trick is that new users must fill in certain data when registering as a new member. Data that can be used in this case is personality trait. Afterward users will be given recommendations that match their personality traits. There are three advantages of using this personality trait [6]. The data pertaining to personality trait can be obtained in two ways, i.e. explicitly (Explicit Personality Elicitation) and implicitly (Implicit Personality Elicitation). The explicit method requires the user to answer a personality trait questionnaire to predict the personality trait. The commonly used personality trait questionnaire is based on the Big Five. As the name implies, Big Five consists of five factors/traits, namely: openness, conscientiousness, extraversion, agreeableness, and intellect. There are many Big Five based questionnaires that are available free of charge ranging from the longest with 504 questions to the shortest with only 10 questions [7].

If a recommender system utilizes the explicit method to obtain a user's personality traits, the user must answer a personality trait questionnaire before becoming a member of the system. Despite the fact that the method can accurately predict a user's personality traits; however, this method is burdensome and time-consuming for the user; therefore, the explicit method is only suitable for use in laboratory studies [8].

To overcome the weaknesses of the explicit personality elicitation method, a researcher may opt to use the implicit personality elicitation method. By using the latter method, a user's personality trait can be predicted, albeit indirectly. The current technique is called the Personality Elicitation from Text (PET). As the name implies, the users' personality traits are predicted from the posts they write, in this case on social media [9][10][11][12][13][14]. However, this method has one obvious weakness when applied to a recommender system, i.e. the user must have at least one active social media account.

In order to cope with the shortcoming of PET, a novel method of implicit personality elicitation is proposed, that is based on demographic data. In this new method, the user's personality is predicted based on demographic data.

To date, demographic data has been applied directly to a recommender system, hence the name Demographic Recommender System. Here the demographic information about the users is used by the classifiers to learn about how to find correlations between certain demographic data with ratings or buying tendencies [15]. However, there has been no research on the use of demographic data to predict human personality traits. The research is useful to overcome the weakness of implicit personality elicitation method, which is based on writing. This is where the gap lies in the personality-based recommender system research, specifically the implicit personality elicitation. This paper presents the result of the work in creating personality model connecting demographic data and personality traits. Next, the model is applied in a personality-based recommender system.

The idea to apply demographic data to predict personality traits come from the results of previous studies which found relationship between personality traits and demographic data. According to [16], one's personality can change or is stable at certain period in the course of his or her life. Except [17], other researchers such as [18] and [19] found that gender also affected personality traits.

Other demographic data such as race/ethnicity/country, hobbies, sport, occupation, zodiac, blood type, and color are known to also affect personality traits. Reference [20] found that persons from different countries have different personality traits. Furthermore, people with different personality traits tend to have different hobbies. An example is a person with a high score in openness is more likely to enjoy something abstract. Therefore, they are most likely are connoisseur of the arts and other forms of culture. Regarding sport, psychologically, a person's preference for a certain type of sport will be supported physically along with preference for a certain movements. Researchers also believe that different personalities will favor different movements.

Researchers also found correlation between personality traits and demographic data such as occupation [21], zodiac [22], blood type [23], and color [24].

The potential benefits of this new method of implicit personality elicitation are that it can be applied to any personality-based recommender system such as in an online shop, library, and travel company. By using this method, the system can give accurate and satisfying recommendations based on the users' demographic data instead of the users' rating history.

A summary of research trends in the recommender system and the proposed method is presented in Fig. 1.

The goal of this research is to create personality and preference models that when applied to a recommender system, then such system:

1) Can be used by any users without the need to have social media account or write status with certain length.

2) Has quite high users' satisfaction.

To achieve the goals, the following research questions must be answered:

1) Which demographic data or combination of demographic data that makes up the best model?

2) When the model is applied to a recommender system, how is the precision of the recommendation and satisfaction to the items recommended?

This paper is structured as follows: the next section, Section II, talk about the methods that are used in processing the data. Results and Discussion is presented in Section III. Section IV concludes the paper.

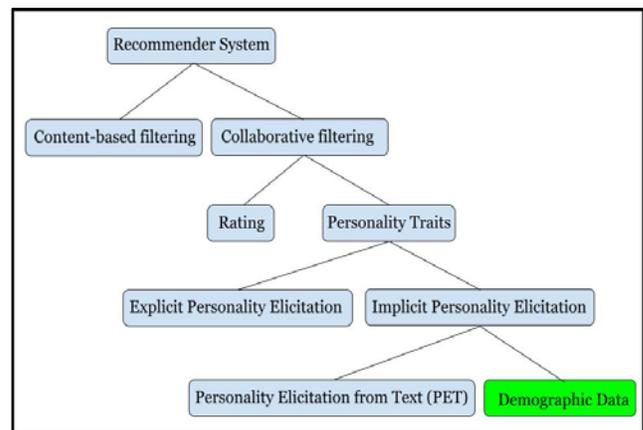


Fig. 1. Research Trends in the Recommender System and the Proposed Method.

## II. METHODS

The detailed survey methods has been presented in [25]. Below is the summary.

A total of 1014 respondents from several cities in Indonesia were involved in the current study. The questionnaire used in this study consists of three parts: i.e. demographic data, personality traits, and preferences.

In this survey, personality questionnaire based on Big Five was used. From a number of Big Five personality trait questionnaires that are available, the Indonesian IPIP 50 questionnaire [26] was chosen. The difference between IPIP 50 and the other Big Five-based questionnaires is that IPIP 50 does not use the term neuroticism; in its place, it uses the term emotional stability which is the opposite of neuroticism. Moreover, IPIP 50 also does not use the term openness; it uses the term intellect instead. To assess the personality traits, respondents were asked to score each question with a score of 1-5 where 1: strongly disagree, 2: disagree, 3: neutral, 4: agree, and 5: strongly agree.

In this survey, the following demographic data were collected: year of birth, marital status, city of residence, sport, occupation, hobbies, ethnicity, favorite color, zodiac, and blood type.

To learn more about preferences, data on respondents' preferences with regard to clothing styles were collected. There

are seven main clothing styles, i.e. Rebellious, Natural, Feminine, Elegant Chic, Dramatic, Creative, and Classic. It should be pointed out that initially this classification of clothing styles was intended for female users; hence, the styles are referred to as feminine and elegant chic. If the terms were applied to a male user, then obviously they would not be referred to as feminine and elegant chic. The term feminine for male users would be substituted with a style that matches the gender, i.e. masculine, whereas, the elegant chic style for male users would simply be referred to as fashionable.

As many as 105 samples of clothes that match the seven clothing styles, or 15 samples for each clothing/fashion style were provided. The respondents were asked to choose samples of clothes they liked. If they liked all the samples then they must choose all and vice versa, if they did not like any of the samples then they did not have to choose any.

After collecting the data, a three steps initial data processing were performed: (1) converted the year of birth into age, (2) calculated the total score of personality traits for each trait, (3) classified the data on the clothes samples selected by the respondents for each clothing style and counted the total number. The number of items selected by the respondent is used to determine the respondent's level of preference for a particular clothing style. If the number of selected items ranges from 1 to 5, then the level of preference is weak. If the number is from 6-10, then the level of preference is moderate. And if the number is greater than that, i.e. between 11 and 15 items, then the level of preference is strong. After that, the three most preferred clothing styles were determined, i.e. the three clothing styles with the highest number of selected items.

While doing the initial processing, the Cronbach alpha value was also calculated to determine the internal consistency of the data. Cronbach alpha was calculated using the following formula:

$$\alpha = \frac{N.c}{v+(N-1).c} \quad (1)$$

where:

$\alpha$  = Cronbach alpha.

N = the number of items.

c = average covariance between item-pairs.

v = average variance.

From the calculation, the following values were obtained: 0.801 for extraversion (good internal consistency), 0.773 for agreeableness (acceptable internal consistency), 0.844 for conscientiousness (good internal consistency), 0.908 for emotional stability (excellent internal consistency), and 0.749 for intellect (acceptable internal consistency).

It should be noted that the questionnaires for the research was made using Google Forms that does not put a limit on how many times a respondent can fill in the questionnaires. Therefore, a check needs to be done to find respondents who fill in the questionnaires more than once. During this check, as many as 25 such respondents were found; therefore, one of the duplicate data was deleted.

In addition to checking for duplicate data, another thing that needs to be checked was the presence of a certain respondent known as a self-enhancer. The presence of a self-enhancer is characterized by a high interscale correlation value, which is the value of the correlation coefficient between attributes. To calculate the interscale correlation, Pearson's correlation coefficient (r) formula was used:

$$r_{xy} = \frac{n(\sum x_i y_i) - (\sum x_i)(\sum y_i)}{\sqrt{[n \sum x_i^2 - (\sum x_i)^2][n \sum y_i^2 - (\sum y_i)^2]}} \quad (2)$$

where:

$r_{xy}$  = correlation between variables x and y.

n = the sample size.

$x_i$  and  $y_i$  = the  $i^{\text{th}}$  sample points.

Since there are five traits, there are 10 correlation coefficients must be calculated such as correlation between extraversion and agreeableness, correlation between extraversion and conscientiousness, etc. All these correlation coefficient values are then averaged to get the average interscale correlation.

The high interscale correlation value can happen because when filling in the questionnaires, self-enhancers tend to rate themselves higher than they should. For that reason, a self-enhancer will show a high level of personality traits in all traits. Therefore, to search for the presence of a self-enhancer, the total scores of personality traits of each respondent must be checked. Respondents who have a maximum score or near to the maximum score on all traits are considered as self-enhancers. In their study, [19] obtained an interscale correlation value of 0.19 and they claimed that such a value indicated that there were not many self-enhancers in the data. However, since a value of 0.38 was obtained, it was assumed that there were quite many self-enhancers in the data. After checking the data, 94 self-enhancers were found. After the data were deleted, the interscale correlation value dropped to 0.24.

When checking the data, data that did not make any sense at all were found. The data were from respondents who rated themselves with a score of 3 on all questions. Therefore, the data were deleted. After deleting the unwanted data, the remaining data is 894.

In building the model, the attribute that act as dependent attribute is level of personality traits. There are two levels: i.e. high and low. The level of personality trait was obtained in the following way: (a) by calculating the average score for each trait, (b) scores that were smaller than averages were designated as low level and scores that were higher than average were labeled as high level.

Additionally, in the modeling, the respondents' age group (categorical type) was used instead of age (numerical type). Therefore, the age data were grouped according to the classification laid down by the Indonesian Ministry of Health [25]. Based on the classification, the respondents were grouped into three groups, i.e. the middle age, adulthood, and adolescents.

The last stage in data processing was to remove some of the attributes that will not be used in the modeling stage. In the modeling, the following attributes were used: blood type, occupation, favorite color, gender, hobby, sport, zodiac group, zodiac component, age group, marital status, ethnicity, intellect level, emotional stability level, conscientiousness level, agreeableness level, extraversion level, preferred clothing style 1, preference level 1, preferred clothing style 2, preference level 2, preferred clothing style 3, and preference level 3.

### III. RESULT AND DISCUSSION

#### A. Modeling

It takes two models to build a personality-based recommender system for fashion; first, a personality model that links the demographic data with personality traits and the other one is a preference model that links the personality traits with a person's preference over fashion. Accordingly, in this stage, the two models were built.

1) *Personality model*: The process to create personality model has been presented in detail in [25]. Below is the summary.

The attributes that were used in the personality trait modeling were blood type, occupation, favorite color, gender, hobby, sport, zodiac group, zodiac component, age group, marital status, ethnicity, intellect level, emotional stability level, conscientiousness level, agreeableness level, and extraversion level. The first eleven attributes are demographic data that serve as independent attributes in the modeling using a decision tree. In addition to using the demographic data individually, a combination of two demographic data (e.g. blood type-occupation, blood type-age group) is used. By combining two demographic data, as many as 54 combinations are obtained, so the total number of demographic data used in the modeling is 65. Hence for each trait, as many as 65 models were created. Meanwhile, the level of personality traits (intellect level, emotional stability level, conscientiousness level, agreeableness level, and extraversion level) were used as the dependent attribute. To evaluate the model, a 10-fold cross validation was used.

Only one model will be used at a later stage. To select the model, the following criteria were used: (1) to make sure that the model can be used by everyone repeatedly at a later date, it has to be made certain that the demographic data in the model will never change, (2) it has to be also made sure that the model are fairly accurate. Based on these criteria, the model

based on age group and gender is chosen [25]. Another reason to choose this model is because previous research found that age and gender has very close relationship to personality traits [19]. Table I presents the model.

2) *Preference model*: As in the personality model, the detailed process in making this preference model has been presented in [27]. The summary is presented.

In the data processing stage, three preferences data and their level (preferred clothing style 1, preference level 1, preferred clothing style 2, preference level 2, preferred clothing style 3, and preference level 3) were selected from each respondent. However, before building the model, a preference data for each clothing style must be created. The data was obtained by combining the respondent's three preferences data into one. Table II shows example preferred data for Natural clothing style that will be used in the modeling. Preferred data for other clothing styles, i.e. Dramatic, Classic, Elegant Chic, Creative, Rebellious, and Feminine were also created.

The data used in building the preference model comprises the levels of the five personality traits and preferences. In the modeling, the class association rule method was used with personality trait's levels as antecedents and preferences as consequent or class.

Association rule is a method to discover a rule that connect items on a transaction. There are at least two measures that are used to identify good rule: support and confidence. If N is the number of transaction, then support of item X, Y is defined as:

$$Support(X) = \frac{Frequency\ of\ X}{N} = P(X) \tag{3}$$

$$Support(X, Y) = \frac{Frequency\ of\ XY}{N} = P(X \cap Y) \tag{4}$$

Meanwhile confidence of  $(X \rightarrow Y)$  is defined as:

$$Confidence(X \rightarrow Y) = \frac{Support(XY)}{Support(X)} = P(Y|X) \tag{5}$$

The personality trait model shown in Table I reveals that the levels of personality traits for extraversion, agreeableness, and conscientiousness are the same for all groups, they are: low for extraversion, high for agreeableness and conscientiousness. Therefore, in this preference modeling, only data with low extraversion, high agreeableness, and high conscientiousness were used. The modeling also only used the data with moderate and strong preference levels. Since it is not possible to recommend men's clothing to women and vice versa; therefore, the men and women were separated.

TABLE I. PERSONALITY MODEL (E: EXTRAVERSION, A: AGREEABLENESS, C: CONSCIENTIOUSNESS, ES: EMOTIONAL STABILITY, I: INTELLECT)

|                    | E   | A    | C    | ES   | I    | Group |
|--------------------|-----|------|------|------|------|-------|
| Adolescence Male   | Low | High | High | Low  | High | 1     |
| Adulthood Male     | Low | High | High | High | Low  | 2     |
| Middle Age Male    | Low | High | High | High | High | 3     |
| Adolescence Female | Low | High | High | Low  | Low  | 4     |
| Adulthood Female   | Low | High | High | High | Low  | 5     |
| Middle Age Female  | Low | High | High | High | Low  | 6     |

TABLE II. SAMPLE OF PREFERRED DATA ON NATURAL CLOTHING STYLE

| Preference   | Level    |
|--------------|----------|
| Elegant Chic | Strong   |
| Natural      | Weak     |
| Natural      | Strong   |
| Natural      | Moderate |
| Natural      | Moderate |
| Rebellious   | Strong   |
| Natural      | Strong   |
| Natural      | Moderate |
| Natural      | Weak     |
| Natural      | Moderate |
| Feminine     | Moderate |
| Natural      | Moderate |

TABLE III. PREFERENCE MODELING RESULT. THE YELLOW CELLS SHOW THE PREFERRED CLOTHING STYLE OF EACH PERSONALITY GROUP (E = EXTRAVERSION, A = AGREEABLENESS, C = CONSCIENTIOUSNESS, ES = EMOTIONAL STABILITY, I = INTELLECT, CL = CLASSIC, CR = CREATIVE, DR = DRAMATIC, EC = ELEGANT CHIC, FE = FEMININE, NA = NATURAL, RE = REBELLIOUS.)

| Group | Personality Traits |      |      |      |      | Fashion Style |      |     |     |     |     |     |
|-------|--------------------|------|------|------|------|---------------|------|-----|-----|-----|-----|-----|
|       | E                  | A    | C    | ES   | I    | Cl            | Cr   | Dr  | EC  | Fe  | Na  | Re  |
| 1     | Low                | High | High | Low  | High | 0,6           | 0,5  | 0,5 | 0,4 | 0,7 | 0,8 | 0,4 |
| 2     | Low                | High | High | High | Low  | 0,5           | 0,7  | 0,7 | 0,8 | 0,5 | 0,8 | 0,7 |
| 3     | Low                | High | High | High | High | 0,6           | 0,26 | 0,3 | 0,7 | 0,6 | 0,9 | 0,3 |
| 4     | Low                | High | High | Low  | Low  | 0,6           | 0,4  | 0,4 | 1   | 0,3 | 0,9 | 0,4 |
| 5     | Low                | High | High | High | Low  | 0,5           | 0,3  | 0,4 | 1   | 0,3 | 0,8 | 0,3 |
| 6     | Low                | High | High | High | Low  | 0,5           | 0,3  | 0,4 | 1   | 0,3 | 0,8 | 0,3 |

TABLE IV. THE PREFERENCE MODEL

| Group | Personality Traits |                    |                        |                          |                | Favorite Fashion Style |
|-------|--------------------|--------------------|------------------------|--------------------------|----------------|------------------------|
| 1     | Low Extraversion   | High Agreeableness | High Conscientiousness | Low Emotional Stability  | High Intellect | Natural, Feminine      |
| 2     | Low Extraversion   | High Agreeableness | High Conscientiousness | High Emotional Stability | Low Intellect  | Elegant Chic, Natural  |
| 3     | Low Extraversion   | High Agreeableness | High Conscientiousness | High Emotional Stability | High Intellect | Natural, Elegant Chic  |
| 4     | Low Extraversion   | High Agreeableness | High Conscientiousness | Low Emotional Stability  | Low Intellect  | Elegant Chic, Natural  |
| 5     | Low Extraversion   | High Agreeableness | High Conscientiousness | High Emotional Stability | Low Intellect  | Elegant Chic, Natural  |
| 6     | Low Extraversion   | High Agreeableness | High Conscientiousness | High Emotional Stability | Low Intellect  | Elegant Chic, Natural  |

Two preferred clothing styles for each personality group are selected based on the highest confidence value. The two fashion styles with the highest confidence value were chosen as the preferred clothing styles (Table III). After selecting the two preferred clothing styles, the preference model can be created and is presented in Table IV.

### B. System Performance

1) *System architecture*: The author in [7] developed a personality-based recommender system in which personality traits were predicted using a questionnaire (explicit method).

Basically, the system consists of two parts, i.e. a part for predicting the personality traits and a part to find the nearest neighbors. Meanwhile, [28] used the personality elicitation from text (implicit method) to predict personality traits in their system. The system they propose basically also consists of two parts, i.e. a part for predicting the personality traits and a part to find the nearest neighbors.

In reference to the two studies above, the recommender system that is built also consists of two parts, i.e. a part for predicting the personality traits and a part to find the nearest neighbors (Fig. 2).

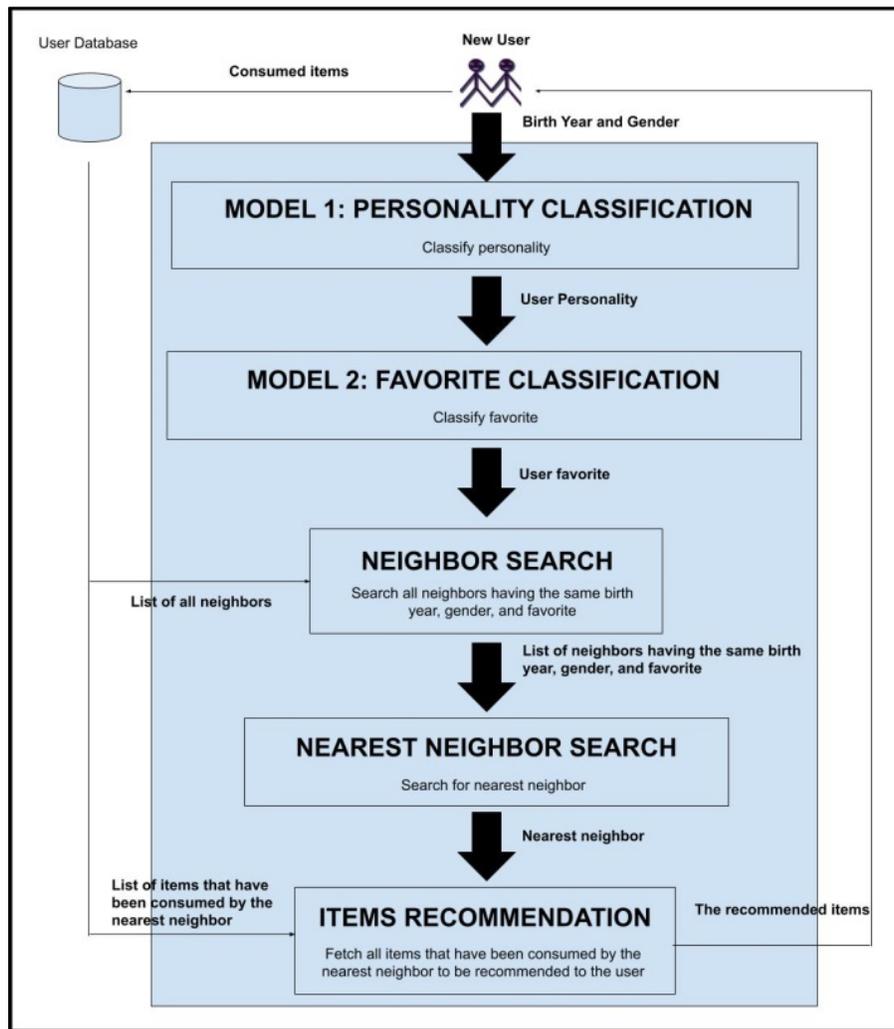


Fig. 2. The Architecture of the Proposed Recommender System.

Before discussing the model, it is necessary to explain what is meant by neighbors—neighbors are respondent's data collected during the data collection stage and modeled at the modeling stage. When the respondents were filling in the questionnaire, they were asked to pick the items that they liked; therefore, this data were treated as data about users who had consumed certain items. These data were stored in the user database as a basis for providing recommendations.

The built system includes a personality model (Model 1) and preference model (Model 2) which was obtained in the previous stage. As mentioned before, the personality model contains rules that link the demographic data to personality traits. Meanwhile, the preference model contains rules that link the personality traits to preferences.

The system starts working when a new user enters the year of birth and gender into the system. Using these data, the system will classify users into certain personality traits (predicting the user's personality traits) based on Model 1 (personality model). These personality traits are passed along to Model 2 (preference model) that will classify user's preferences based on the user's personality traits (predicting user's preferences).

After generating user's preferences, the system will retrieve all neighbors in the user's database in search of all neighbors whose year of birth, gender, and preferences are the same as the user's year of birth, gender, and preferences. The result is a list of neighbors with the same year of birth, gender, and preferences as the user.

By using the filtered list of neighbors, the system will search for the nearest neighbor. It should be explained that in this study, all users are using the model for the first time. Therefore, the user has never consumed any items. Because of that, at this stage, the system looks for the nearest neighbor based only on the same year of birth. If there are several neighbors with the same year of birth, then the system will pick one neighbor.

After the nearest neighbor is obtained, the system will collect all the items that have been consumed by the nearest neighbor. These are the items that will be recommended to the users. After consuming some or all of the recommended items, the new user data will be saved to the user database.

2) *System evaluation:* A number of researchers have used direct method to evaluate the systems they have built. A survey

conducted by [29] involved 21 students as the respondents. In the survey, the respondents were asked about the novelty of the items recommended, the accuracy, and satisfaction with the recommendations given. In another study, [30] also carried out a direct survey to users to find out the respondents' level of satisfaction with the built model.

To evaluate the performance of the proposed recommender system, a direct survey involving 74 respondents was conducted. In the survey, the respondents were asked to interact with the system.

- Relevance of the recommended items. In this experiment, the system recommends a number of items to the user. The system provides a Like button on all recommended items. When checking the recommended items one by one, if the button was pressed, it means that the item is relevant. Then the precision of the relevant items can be calculated. Precision is the percentage of relevant items from all the recommended items.

$$\text{Precision} = \frac{\text{Number of relevant items}}{\text{Number of recommended items}} \quad (6)$$

Here each user has one precision value, whereas the system's precision is the average precision of all users.

- Satisfaction to the recommended items. To find out user's satisfaction, the CSAT method is used. By using the method, user's satisfaction can be gauged by asking them how satisfied they are with the goods or services they used. There are five scales used to assess user's satisfaction, i.e. very dissatisfied, dissatisfied, neutral, satisfied, and very satisfied. The percentage of satisfaction is calculated using the following formula:

$$\text{Satisfaction} = \frac{\text{Number of responses}}{\text{Total number of responses}} \times 100\% \quad (7)$$

Note that the responses used in the formula are only Satisfied and Very Satisfied ones.

In the study, the system will ask the question: "How do you rate the recommended items?" immediately after the respondent finished checking the recommended items. The respondents were supplied with five answers as follows: very dissatisfied, dissatisfied, neutral, satisfied, and very satisfied. The percentage of user satisfaction is only calculated from the responses that give the value of Satisfied and Very Satisfied.

### C. Evaluation Result

From the evaluation, the following facts were obtained:

- 1) Precision of the recommendation was 60.19%.
- 2) Satisfaction to the items recommended was 87.50%.
- 3) There were as many as 17 respondents whose precision less than 50%, nevertheless satisfied or even very satisfied with the recommendation.
- 4) There were as many as 5 respondents whose precision more than 50%, yet dissatisfied (neutral) with the recommendation.

From the facts above, it can be said that a respondent with low precision can still be satisfied with the recommendation. In opposite, a respondent with high precision can be dissatisfied with the recommendation. Therefore, it is hypothesized that satisfaction correlates with level of preference, and not with precision. The statement on satisfaction has no correlation with accuracy has been confirmed by [31] and [32]. Note that precision is one of accuracy metrics used in recommender system besides recall, MAE, and MSE.

User's satisfaction is a psychological condition that can be measured from the user's expectation. A user will satisfy if the products or services offered to them exceeds or at least the same as the expectation. On the contrary, a user does not satisfy if the experience when using the products or services below the expectation [33]. Based on this, the reason why 17 respondents with low precision still satisfied or very satisfied with the recommendation is because they like the items recommended very much (high preference level) despite they only like a few of the many items recommended. In other word, the user's expectations are fulfilled. Meanwhile, in the case of 5 respondents with high precision but dissatisfy; it is because they do not really like the items recommended (medium preference level). As a result, they dissatisfy with the recommendation although they like many items. In other word, the user's expectation still not met.

The author in [29] reported that the precision and satisfaction of PET-based system were 82% and 79%, respectively. Compared to demographic data-based system performance, the precision of PET-based system is better but demographic data-based system is better in satisfaction. In recommendation system, accuracy is important but accuracy alone is not enough. This is because user satisfaction is more important. The next paragraph explains about this.

According to [31] two main tasks of accuracy metrics in recommender system are:

- 1) To measure the accuracy of single prediction. This is called predictive accuracy metrics. This metric calculate how close the predicted rating from the actual rating. Mean absolute error (MAE) and mean squared error (MSE) are used in this metric.
- 2) To evaluate the effectivity of the system in selecting the high quality items from a set of available items. This metric is called decision-support metrics and uses precision and recall to declare the accuracy.

Furthermore, [31] stated that building a recommender system with high accuracy was not enough. This is because the most accurate recommendation based on those metrics above, is sometimes not the useful recommendation for or liked by the users. This causes dissatisfaction to the users. In other word, user satisfaction does not always correlate with high accuracy [32][31]. Knowing the importance of user's satisfaction, [31] and [34] stated that it is not fair to judge a recommender solely from its accuracy. The users must be taken into account since they do not care with the algorithm to increase the recommendation accuracy. They only want to have useful recommendations.

Back to the performance comparison between demographic data-based system and PET-based one where the demographic data-based system has lower accuracy but higher satisfaction. Based on the above explanation, recommender system whose satisfaction is higher is better. In other word, demographic data-based system is better than the PET-based system.

#### IV. CONCLUSION

A research has been carried out to find model connecting demographic data and personality traits. As many as 65 models of each trait were created. From those models, the one based on age group and gender was selected as the working model since it satisfied two criteria. Besides that, previous research also found that age and gender had very close relationship to personality traits.

From the performance evaluation, precision and satisfaction of the demographic data-based recommender system were 60.19% and 87.50% respectively. When compared to PET-based system, demographic data-based system is lower in precision but higher in satisfaction. Other advantage of demographic data-based system compared to PET-based system is there is no obligation to have social media account.

Despite the strength of demographic data-based system compared to PET-based system, this research has some limitations:

- 1) The fashion classification may differ from other classifications. So, when a clothing is classified into a certain clothing style, others may classify it into different clothing style.
- 2) Some of the respondents may be come from low-income communities who do not care with fashion hence the choice of preferred clothing different from other respondents with the same category (e.g. age group – gender).

To test the hypothesis about the relationship between satisfaction and level of preference, another experiment is needed. In the experiment the respondents are asked not only to determine whether they like an item or not, but also the level of preference to the item. One way to obtain the level of preference is by providing five stars on each item. The more stars given by a respondent to an item, the higher the level of preference of the respondent to that item.

#### REFERENCES

[1] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating 'word of mouth,'" *Conf. Hum. Factors Comput. Syst. - Proc.*, vol. 1, pp. 210–217, 1995.

[2] H. Li, F. Cai, and Z. Liao, "Content-based filtering recommendation algorithm using HMM," *Proc. - 4th Int. Conf. Comput. Inf. Sci. ICCIS 2012*, no. March, pp. 275–277, 2012, doi: 10.1109/ICCIS.2012.112.

[3] R. Burke, A. Felfernig, and M. H. Göker, "Recommender systems: An overview," *Assoc. Adv. Artif. Intell.*, 2011.

[4] T. Bogers and A. van den Bosch, "Collaborative and Content-based Filtering for Item Recommendation on Social Bookmarking Websites," in *Proceedings of the ACM RecSys'09 Workshop on Recommender System & the Social Web*, 2009, doi: 10.1007/978-3-642-25694-3.

[5] N. Rubbens, M. Elahi, M. Sugiyama, and D. Kaplan, "Active Learning in Recommender Systems," in *Recommender Systems Handbook*, Second Edi., F. Ricci, L. Rokach, and B. Shapira, Eds. New York: Springer, 2015, pp. 809–846.

[6] I. Paryudi, A. Ashari, and A. M. Tjoa, "Personality Estimation using

Demographic Data in a Personality-based Recommender System: A Proposal," 2019, doi: 10.1145/3366030.3366098.

[7] M. Augusta Silveira Netto Nunes, "Recommender Systems based on Personality Traits," *Université Montpellier II*, 2008.

[8] M. Tkalčić and L. Chen, "Personality and Recommender Systems," in *Recommender Systems Handbook*, Second Edi., F. Ricci, L. Rokach, and B. Shapira, Eds. New York: Springer, 2015, pp. 715–739.

[9] V. Ong, A. D. S. Rahmanto, W. Williem, N. H. Jeremy, D. Suhartono, and E. W. Andangsari, "Personality Modelling of Indonesian Twitter Users with XGBoost Based on the Five Factor Model," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, pp. 248–261, 2021, doi: 10.22266/ijies2021.0430.22.

[10] A. V. Kunte and S. Panicker, "Using textual data for Personality Prediction:A Machine Learning Approach," in *2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019*, 2019, no. March, pp. 529–533, doi: 10.1109/ISCON47742.2019.9036220.

[11] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *J. Big Data*, vol. 8, no. 1, pp. 1–20, 2021, doi: 10.1186/s40537-021-00459-1.

[12] A.-R. Feizi-Derakhshi, M.-R. Feizi-Derakhshi, M. Ramezani, N. Nikzad-Khasmakhi, and M. Asgari-Chenaghlu, "The state-of-the-art in text-based automatic personality prediction," *arXiv*, pp. 1–27, 2021.

[13] Z. Wang, C. H. Wu, Q. B. Li, B. Yan, and K. F. Zheng, "Encoding text information with graph convolutional networks for personality recognition," *Appl. Sci.*, vol. 10, no. 12, 2020, doi: 10.3390/APP10124081.

[14] X. Wang, Y. Sui, K. Zheng, Y. Shi, and S. Cao, "Personality classification of social users based on feature fusion," *Sensors*, vol. 21, no. 20, 2021, doi: 10.3390/s21206758.

[15] C. C. Aggarwal, *Recommender Systems the Textbook*, First Edit. New York: Springer, 2016.

[16] M. A. Harris, C. E. Brett, W. Johnson, and I. J. Deary, "Personality Stability From Age 14 to Age 77 Years," *Psychol. Aging*, vol. 31, no. 8, pp. 862–874, 2016, doi: 10.1037/pag0000133.supp.

[17] J. S. Hyde, "The gender similarities hypothesis," *Am. Psychol.*, vol. 60, no. 6, pp. 581–592, 2005, doi: 10.1037/0003-066X.60.6.581.

[18] M. Vianello, K. Schnabel, N. Sriram, and B. A. Nosek, "Gender Differences in Implicit and Explicit Personality Traits," *Pers. Individ. Dif.*, 2013, doi: 10.2139/ssrn.2249080.

[19] C. J. Soto, O. P. John, S. D. Gosling, and J. Potter, "Age Differences in Personality Traits From 10 to 65: Big Five Domains and Facets in a Large Cross-Sectional Sample," *J. Pers. Soc. Psychol.*, vol. 100, no. 2, pp. 330–348, Feb. 2011, doi: 10.1037/a0021717.

[20] D. P. Schmitt et al., "The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations," *J. Cross. Cult. Psychol.*, vol. 38, no. 2, pp. 173–212, 2007, doi: 10.1177/0022022106297299.

[21] D. Heller, W. Q. E. Perunovic, and D. Reichman, "The future of person-situation integration in the interface between traits and goals: A bottom-up framework," *J. Res. Pers.*, vol. 43, no. 2, pp. 171–178, 2009, doi: 10.1016/j.jrp.2008.12.011.

[22] J. J. F. van Rooij, "Introversion-extraversion: astrology versus psychology," *Pers. Individ. Dif.*, vol. 16, no. 6, pp. 985–988, 1994, doi: https://doi.org/10.1016/0191-8869(94)90243-7.

[23] M. Rogers and A. I. Glendon, "Blood type and personality," *Pers. Individ. Dif.*, vol. 34, pp. 1099–1112, 2003.

[24] R. Miao, "Colour preference," *Birmingham*, 2017.

[25] I. Paryudi, A. Ashari, and K. Mustofa, "Modelling Relationship between Demographic Data and Personality Traits," *Turkish J. Comput. Mat. Educ.*, vol. 12, no. 14, pp. 2247–2255, 2021.

[26] H. Akhtar and S. Azwar, "Indonesian Adaptation and Psychometric Properties Evaluation of the Big Five Personality Inventory: IPIP-BFM-50," *J. Psikol.*, vol. 46, no. 1, pp. 32–44, 2019, doi: 10.22146/jpsi.33571.

[27] I. Paryudi, A. Ashari, and K. Mustofa, "Creating Personality and Preference Models based on Demographic Data for Personality-based Recommender System for Fashion using Decision Tree and Association

- Rule,” Turkish J. Comput. Mat. Educ., vol. 12, no. 14, pp. 5165–5174, 2021, [Online]. Available: <https://turcomat.org/index.php/turkbilmater/article/view/11548>.
- [28] A. Roshchina, J. Cardiff, and P. Rosso, “TWIN: Personality-based Intelligent Recommender System,” J. Intell. Fuzzy Syst., vol. 28, no. 5, pp. 2059–2071, 2015, doi: 10.3233/IFS-141484.
- [29] A. Di Rienzo and A. Neishabouri, “Recommendations with personality traits extracted from text reviews,” in Studies in Computational Intelligence, vol. 616, Springer Verlag, 2016, pp. 355–364.
- [30] R. Hu, “Design and User Issues in Personality-based Recommender Systems,” in RecSys’10, 2010, p. 386.
- [31] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, “Improving recommendation lists through neighbor diversification,” 2005.
- [32] S. M. McNee et al., “On the Recommending of Citations for Research Papers On the Recommending of Citations for Research Papers,” in Proc. of ACM CSCW 2002, 2002, no. November 2002, doi: 10.1145/587078.587096.
- [33] F. Tjiptono, Service, Quality, and Satisfaction. Yogyakarta: Penerbit Andi, 2016.
- [34] C. Hayes, P. Massa, P. Avesani, and P. Cunningham, “An on-line evaluation framework for recommender systems,” in Workshop on Personalization and Recommendation in E-Commerce, 2002, no. May.

# A Greedy-based Algorithm in Optimizing Student's Recommended Timetable Generator with Semester Planner

Khyrina Airin Fariza Abu Samah<sup>1</sup>, Siti Qamalia Thusree<sup>2</sup>

Ahmad Firdaus Ahmad Fadzil<sup>3</sup>, Lala Septem Riza<sup>4</sup>, Shafaf Ibrahim<sup>5</sup>, Noraini Hasan<sup>6</sup>

Faculty of Computer and Mathematical Sciences<sup>1,2,5,6</sup>

Universiti Teknologi MARA Cawangan Melaka Kampus Jasin, Melaka, Malaysia<sup>1,2,5,6</sup>

Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Shah Alam, Selangor, Malaysia<sup>3</sup>

Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia<sup>4</sup>

**Abstract**—Semester planner plays an essential role in students' society that might help students have self-discipline and determination to complete their studies. However, during the COVID-19 pandemic, they faced difficulty organizing time management and doing a manual schedule. It resulted in substantial disruptions in learning, internal assessment disturbances, and the cancellation of public evaluations. Hence, this research aims to optimize the recommended semester planner, Timetable Generator using a greedy algorithm to increase student productivity. We identified three-set control functions for each entered information: 1) validation for the inserted information to ensure valid data and no redundancy, 2) focus scale, and 3) the number of hours to finish the activity. We calculate the priority task sequence to achieve the best optimal solution. The greedy algorithm can solve the optimization problem with the best optimal solution for each situation. Then, we executed it to make a recommended semester planner. From the test conducted, the functionality shows all the features successfully passed. We validate using test accuracy for the system's reliability by evaluating it compared to the Brute Force algorithm, and the trends increase from 60% to 100%.

**Keywords**—Greedy algorithm; optimization; recommendation system; semester planner

## I. INTRODUCTION

The COVID-19 pandemic has made a huge impact on the educational system around the world, forcing lockdown in schools, universities, and colleges. Hence, Singh and Kant [1] claimed it caused a significant disrupted student learning, internal examinations, and the termination of public exams for qualification due to inferior alternatives. Based on the research, most students are not well prepared for classes. The condensed class schedule negatively impacts their learning experiences [2]. It corresponded with low experiences and academic achievement by the student.

A survey conducted towards 126 students of Universiti Teknologi MARA Melaka Kampus Jasin. They responded that they used paper and pen to organize semester planners manually, and 84.9% of students found it challenging to organize semester planners with upcoming quizzes and tests.

Additional time to new learning content is taught to enhance their performance. However, the effect of this treatment is a slight difference throughout the distribution of student performance. While the most benefit is for high-performing students, low-performing students do not benefit. Foulkes et al. [3] highlight that the research argued that it is crucial to determine the content of additional time instruction to explain this pattern.

Besides, balancing studying with the fun stuff in life can be challenging to maintain. Unstructured academic lifestyle is one of the impacts of poor sleep quality in university life [4]. It increases mental the risk of mental illness and poor academic achievement. Panda et al. [5] mentioned that based on research on academic stress students, the stress factors include many assignments, competition with other students, failures, financial problems, poor relationships, lectures, or family matters.

Thus, planning a scheduler is an effective solution to organize and navigate students' time management [6]. A semester planner is an organized schedule where students can create outlines and awareness of the study methods and the effectiveness of their study practices for the current semester and learning goals. Students become self-discipline, deterministic, and more confident [7]. In addition, it helps them build self-regulated learning skills to succeed in various learning environments and workplaces. As a primary time management tool, a scheduler comes in a list of times when events and actions occur. Brioso et al. [8] supported that planning and controlling did not starve due to insufficient input in an intrinsically challenging task.

The primary substance of time management is reducing stress [9]. Of course, building a schedule for the current semester is a tedious task. Students need to fit all the courses into a single scheduler, which would take time and patience. The time management principle is time management on how to plan the specific tasks and work and organized. Otherwise, it can affect entire planning and work. Based on the same survey, 83.8% of the students claimed they never use revision planners' automated class timetable generator.

The research was sponsored by Universiti Teknologi MARA Cawangan Melaka under the TEJA Grant 2021 (GDT 2021/1-28)

Therefore, to overcome the problem, a recommended semester planner using optimization of the greedy-based algorithm was used quickly to find an approximate solution in the optimization prickly. This parallel with the greedy-based algorithm concept, to solve problems in getting the globally optimal solution with the choice seems best by making a locally optimal choice [10]. Moreover, the greedy allocation is compatible with a dynamic heterogeneous resource environment linked to the scheduler via a homogenous communication environment [11]. Thus, Choudhary and Peddoju [12] claimed that it is easier to assess the run with more efficient use of available resources. It allows UiTM students to identify peak study load times to ensure they have plenty of time to complete all assignments. The achievement of each target would bring in the motivation to achieve another target.

## II. RELATED WORK AND TECHNIQUES

This section describes the related works in scheduling and related techniques related to the study.

### A. Related Work in Scheduling

There are some related works in scheduling such as course timetables, high school timetables [13], exam timetables, and some of the other schedules used in the workplace. Although many intelligent practices have been used in recent years to solve course timetabling, there are still many areas for improvement. Wang et al. [13] claimed that to effectively solve the Greedy Algorithm and provide a high-performance initial Genetic Algorithm population, they suggested using a Greedy and Genetic Fusion Algorithm. The hybrid Cat Swarm Optimization algorithm-based application solves the school schedule problem. It is efficient, easy to use, and fast. The demonstration conducts by experiments with real-life input data to test the efficiency and performance. The test case uses the same timetable and the same description statistics. The result shows that the hybrid CSO-based algorithm performs better in less computational time than many other existing approaches.

For timetable examination, a new proposed combination of three approaches: multiple criteria, maximum independent set, and heuristic graph to solve the problem of timetable examination [14]. There is also have some related work on schedulers in the workplace, such as flight schedules. Optimization control and algorithm design have also been hot issues in China's research into unusual flights. Based on a better knowledge of the NP problem under the flight recovery model, the researcher validates the usefulness of the greedy random adaptive algorithm in the problem-solving process [15].

### B. Related Techniques

An algorithm is a step-by-step approach to problem-solving, widely used for data processing, calculation, and other related mathematical and computer operations. Different algorithms can easily and quickly perform operations or solve problems in terms of efficiency. There are three related techniques: Evolutionary Algorithm (EA), Greedy Algorithm, and Brute Force.

The Evolutionary Algorithm (EA) development has been significant among the research and optimization techniques set in the last decade [16]. EAs are a set of modern heuristics used with great complexity in many applications. It was the engine of a field known as Evolutionary Computation (EC) that successfully solved complex problems [17]. The advantages of EC techniques derived from gaining fitness and flexibility to the goal combined with robust behavior. Nowadays, EC is considered an adaptable problem-solving concept, particularly in complex optimization problems.

A greedy-based algorithm produces a good solution to some math but not others. As mentioned by Prabhakaran et al. [18], most of the issues they are working on will have two properties; firstly, a greedy choice property is defined as the best way to make any choice to solve sub-problems that might arise later. A greedy algorithm's choice might depend on the decision but not the future or all the subproblem's solutions. It makes iteratively one greedy choice to minimize a problem into an optimal solution. Secondly, an optimal substructure is an optimal solution to solve the problem includes optimal solutions to the subproblem. Then a will problem displays optimal substructure. The problem of optimization is finding the best solution that seems all feasible [19].

Brute Force defines a straightforward problem strategy, commonly dependent on problem statement and concept definition. Definition of 'force' is a computer strategy but not intellect. Another approach to describe Brute Force is 'Just do it', and sometimes, the technique is easiest to use [20]. Exhaustive or Brute Force searches can generate and test the basic solving problem method and algorithm paradigm. Mahoor et al. [21] claimed it consists of the listed systematic potential candidate for solution and testing if each candidate meets the problem statement. Therefore, this approach is used commonly when the problem is small enough that heuristics can reduce the number of possible solutions to a manageable number. Also, identifying a solution recovery model takes more time than speed.

As a result, an evolutionary algorithm is beneficial with conceptual simplicity, while Brute Force is widely applicable and known for its simplicity. However, the greedy-based algorithm is more accessible to implement and faster than the other two. Although the Greedy-based algorithm has not always reached the optimal global solution, most of the time reaches the sub-optimal solution. Thus, greedy-based techniques are the best solution for this project in schedule recommendation of timetable generator with semester planner.

## III. METHODS

This section divides research methodology into four phases: gather information, design and development and system validation.

### A. Gather Information Phase

The survey was conducted on 126 respondents of students from UiTM Cawangan Melaka Kampus Jasin on their experience to understand better the real problem faced by them. Table I shows the summary of survey information. About 85.70% of respondents agreed that they have difficulties organizing their study planner, while the rest, 14.30%, voted

with no issues. Some of them never use study planners with 58.13%, followed by book or handphone planners with 24.78%, sticky notes with 14.60%, and timetable 2.49%.

TABLE I. SURVEY DETAILS AND RESPONSES TO GATHER INFORMATION

| Questions                              | Item Response          | Percentage Response |
|----------------------------------------|------------------------|---------------------|
| Difficulties To Organize Study Planner | Yes                    | 85.70%              |
|                                        | No                     | 14.30%              |
| Current Method Study Planner           | Never                  | 58.13%              |
|                                        | Book/Handphone Planner | 24.78%              |
|                                        | Use Sticky Note        | 14.60%              |
|                                        | Timetable              | 2.49%               |

**B. Timetable Generator Flowchart**

We use the flowchart as in Fig. 1 to design how the timetable generator operates by defining the system’s module, data, and architecture to specify the system requirements. It starts with key-in details information for the user’s hobby, extracurricular, and subject details for the current semester. We set the control functions for each information: 1) validation for the inserted information to ensure valid data and no redundancy, 2) focus scale, and 3) the number of hours to finish the activity. The information entered by the user can be updated or deleted before we generate the planner using the Greedy algorithm.

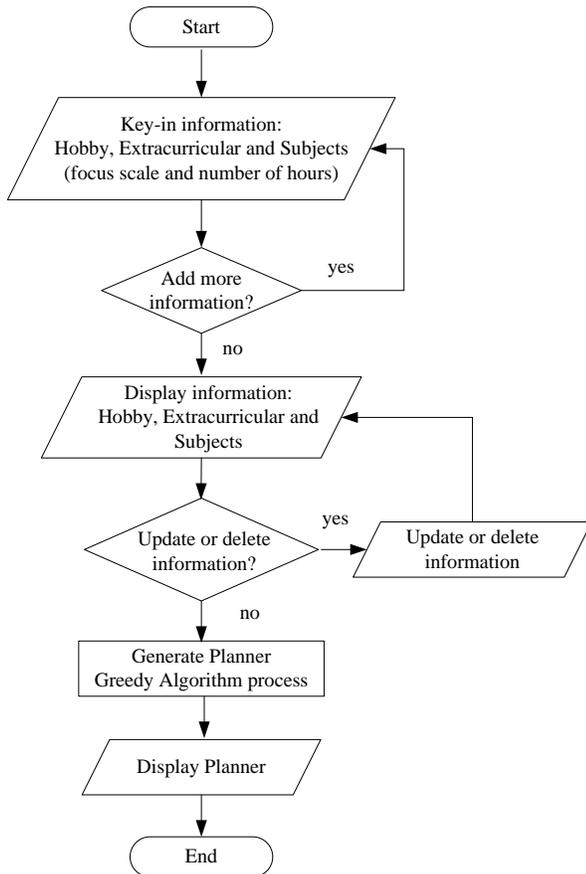


Fig. 1. Flowchart for Timetable Generator Planner.

**C. System Interface**

System design or user interface design is designing an interface in software or computerized devices focusing on looks or style. This research aims to design an interface that is easy to use and understandable by the user. User interface design usually refers to graphical user interfaces.

**D. Greedy Algorithm Development**

Fig. 2 shows the Greedy algorithm flowchart of the phase involved in this algorithm to achieve the best optimization result. The algorithm needs to solve scheduling problems in tasks that need to be completed, required time to complete each task, and priority of each task. Three main processes involved:

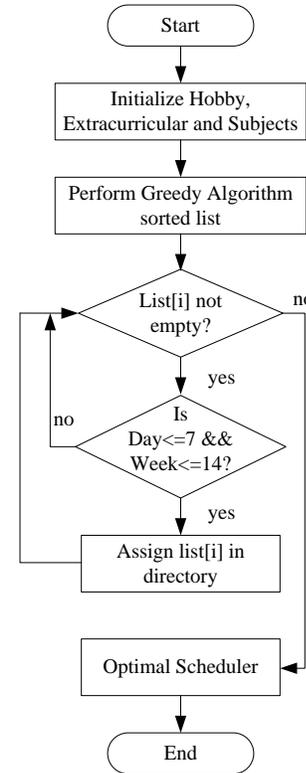


Fig. 2. Phases in Optimizing Greedy Algorithm for Timetable Generator.

1) *Step 1: Initialization:* The algorithm initializes all user variables and sorts the list with the greedy approach. The data will be classified into three lists: name, focus scale, and complete time.

2) *Step 2: Sort list:* Sort the tasks based on decreasing order of priority. However, in some cases, the priorities of different tasks are the same. This algorithm recommends the list of tasks that require a higher focus scale and the shortest time completed the task to minimize the problem. The algorithm would give preference to a higher focus scale and tasks that take less time to complete so that the higher focus scale will lead to a higher score, and more time needed will reduce the score. After being classified into a list, this phase has created another list of pairs [key, value] as shown below. Each of the list pairs is a combination of hobby, extracurricular, and subject.

- Complete time [complete time, name].
- Focus scale [focus scale, name].
- Priority [(focus scale/complete time), name].

Then, the system would proceed with the sorting process for each list in decreasing the order of “key”. After all the variables have been sorted, the algorithm will assign lists in a directory with a specific condition.

3) *Step 3: Optimal scheduler:* Assume that task in priority sequence is  $n$ , set priority = {(8, ICT662), (7, Badminton), ... $n$ } and tasks use the same resources from of hobby, extracurricular and subject task, with priority score for each task. The algorithm use planner set to store the selected task. The task chosen is ICT662, and Monday is initialized “ICT662”. Then the algorithm will go through 14 weeks to check whether the task has an assignment, quiz, or test week. For example, ICT662 has a quiz on week five, then Monday on week five is initialized “Quiz 1 ICT662”. An algorithm will proceed initialized in planner set with next day and weeks until all the task in the priority list is empty.

#### E. System Validation

Lastly, we proceed with system validation to test the functionality and reliability of the system. A functionality test runs to test whether the system’s function is running smoothly or not. If there are bugs when running the functionality test, we need to correct the function to ensure it meets the requirements. Functionality testing tests and identifies which function is performed with or without error. If an error occurs in this phase, we need to fix the error before the implementation. For the reliability test, we compare the result of the Timetable Generator with the Brute Force algorithm to check the result’s accuracy. We tested this testing phase with developers, lectures, and UiTM Jasin students to know whether this project is performing well or not.

### IV. RESULT AND DISCUSSION

#### A. Functionality Testing

We have to test and check the core’s function application, input, button, table, and more during this phase. To test the system’s functionality, Table II describes consideration of

features criterion where we test it accordingly and fill up the result. There are six features involved: registration and login, subject, hobby and extracurricular, updating user’s details, a sequence of the focus scale, complete time and priority, planner and print planner.

The output of Timetable Generator, as in Fig. 3, display the sorted list based on the focus scale, complete time and priority in the list box. These list boxes show the pair between focus scale and complete time with all user tasks. The sorted priority list box is the greedy algorithm result sequence with the score and name of each task. Finally, Fig. 4 shows the output result of the Timetable Generator or the semester planner with an optimal scheduler. We successfully passed all the features, and finally, users can print a semester planner in pdf form as the last feature development.

#### B. Reliability Testing

We validate the reliability testing based on the system’s accuracy in this phase and use ten tests to run the validation. The system’s response might have a different condition with different sequences of priority and Brute Force result on the hobby, extracurricular, and subject enrol. Table III shows the output from the system prototype and Brute Force algorithm result by manual calculation. We test the accuracy testing based on these two rules: First, give preference to higher priorities, leading to a higher score. Second, give preference to tasks that take less time to complete, so tasks that take more time will reduce the score. Next, we calculate the accuracy based on the similarity result sequence of Brute Force and Timetable Generator.

Table IV shows the summary comparison on the research accuracy result,  $p$  between Brute Force and our proposed study. Only attempt 1 has a slight difference out of five; only three items have similar according to the result sequence. The rest outcome gained is similar for both reliability tests.

We constructed the visualization of an accuracy graph from the result of 10 output. Fig. 5 shows the 2-axes graph, increasing from 0.60 (60%) to 1.0 (100%). It proves that the Timetable Generator system is accurate and can be used for the student to overcome the problem.

TABLE II. FEATURES COMPARISON BETWEEN RECOMMENDATION TECHNIQUES

| Features                                              | Details                                                                                                                                                                                                                                                                                                                | Status |
|-------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| Registration and login                                | The system can register a new account with an encrypted password and allow users to login into the system with the correct combination of attributes needed.                                                                                                                                                           | Passed |
| Subject, hobby and extracurricular                    | The system can store the user’s subject, hobby, and extracurricular details in the database. Before storing the details, the system will check whether all of the data are inserted or not. The incomplete information would not allow being stored in the database.                                                   | Passed |
| Updating user’s details                               | The system retrieved all user details from the database and displayed them in a grid view. Users can update some data and store it back in the database.                                                                                                                                                               | Passed |
| A sequence of complete time, focus scale and priority | The system can list decreasing order of complete time, focus scale and priority.                                                                                                                                                                                                                                       | Passed |
| Planner                                               | The system prescribes a schedule of semester planners that is the best based on the priority sequence. The schedule will be displayed in the form of a table that contains seven days and 14 weeks of a semester. Thus, the system will display the subject details based on assignment week, quiz week and test week. | Passed |
| Print planner                                         | Print the planner in the form of a PDF.                                                                                                                                                                                                                                                                                | Passed |

| Focus Scale                                                                                                                                                                 | Complete Time                                                                                                                                                               | Priority                                                                                                                                                                                      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [9, DCS651]<br>[9, CSC649]<br>[9, CSC580]<br>[8, ICT663]<br>[4, Jawatankuasa Pengurusan Kolej]<br>[3, TKC501]<br>[3, Futsal]<br>[2, CSC253]<br>[2, Netball]<br>[2, Fishing] | [5, DCS651]<br>[3, Netball]<br>[3, Fishing]<br>[2, TKC501]<br>[2, CSC253]<br>[2, ICT663]<br>[2, CSC649]<br>[2, CSC580]<br>[2, Jawatankuasa Pengurusan Kolej]<br>[2, Futsal] | [4.5, CSC649]<br>[4.5, CSC580]<br>[4, ICT663]<br>[2, Jawatankuasa Pengurusan Kolej]<br>[1.8, DCS651]<br>[1.5, TKC501]<br>[1.5, Futsal]<br>[1, CSC253]<br>[0.667, Netball]<br>[0.667, Fishing] |

Fig. 3. Snapshot Output of a Sorted List of Information in Timetable Generator.

| Day       | Week 1                               | Week 2                               | Week 3                               | Week 4                                                              | Week 5                                                                | Week 6                                                          | Week 7                                                 | Week 8                                                             | Week 9                                           | Week 10                                   | Week 11                                         | Week 12                                   | Week 13                                    | Week 14                              |
|-----------|--------------------------------------|--------------------------------------|--------------------------------------|---------------------------------------------------------------------|-----------------------------------------------------------------------|-----------------------------------------------------------------|--------------------------------------------------------|--------------------------------------------------------------------|--------------------------------------------------|-------------------------------------------|-------------------------------------------------|-------------------------------------------|--------------------------------------------|--------------------------------------|
| Monday    | ~CSC649,<br>~CSC253                  | ~CSC649,<br>~CSC253                  | ~CSC649,<br>~CSC253                  | ~CSC649,<br>~CSC253                                                 | ~CSC649,<br>~Test 1<br>CSC649,<br>~CSC253,<br>~Assignment 1<br>CSC253 | ~CSC649,<br>~Quiz 1<br>CSC649,<br>~CSC253,<br>~Quiz 1<br>CSC253 | ~CSC649,<br>~Assignment 1<br>CSC649,<br>~CSC253        | ~CSC649,<br>~CSC253,<br>~Assignment 2<br>CSC253, ~Test<br>1 CSC253 | ~CSC649,<br>~CSC253                              | ~CSC649,<br>~CSC253,<br>~Quiz 2<br>CSC253 | ~CSC649,<br>~CSC253,<br>~Assignment 3<br>CSC253 | ~CSC649,<br>~Test 2<br>CSC649,<br>~CSC253 | ~CSC649,<br>~Test 2<br>CSC253              | ~CSC649,<br>~CSC253                  |
| Tuesday   | ~CSC580,<br>~Netball                 | ~CSC580,<br>~Netball                 | ~CSC580,<br>~Netball                 | ~CSC580,<br>~Assignment 1<br>CSC580, ~Quiz<br>1 CSC580,<br>~Netball | ~CSC580,<br>~Test 1<br>CSC580,<br>~Netball                            | ~CSC580,<br>~Netball                                            | ~CSC580,<br>~Quiz 2<br>CSC580,<br>~Netball             | ~CSC580,<br>~Assignment 2<br>CSC580,<br>~Netball                   | ~CSC580,<br>~Netball                             | ~CSC580,<br>~Netball                      | ~CSC580,<br>~Netball                            | ~CSC580,<br>~Netball                      | ~CSC580,<br>~Test 2<br>CSC580,<br>~Netball | ~CSC580,<br>~Netball                 |
| Wednesday | ~ICT663,<br>~Fishing                 | ~ICT663,<br>~Fishing                 | ~ICT663,<br>~Fishing                 | ~ICT663,<br>~Fishing                                                | ~ICT663,<br>~Fishing                                                  | ~ICT663,<br>~Fishing                                            | ~ICT663, ~Quiz<br>1 ICT663,<br>~Fishing                | ~ICT663, ~Test<br>1 ICT663,<br>~Fishing                            | ~ICT663,<br>~Assignment 1<br>ICT663,<br>~Fishing | ~ICT663,<br>~Fishing                      | ~ICT663,<br>~Fishing                            | ~ICT663,<br>~Fishing                      | ~ICT663,<br>~Fishing                       | ~ICT663,<br>~Fishing                 |
| Thursday  | ~Jawatankuasa<br>Pengurusan<br>Kolej | ~Jawatankuasa<br>Pengurusan<br>Kolej | ~Jawatankuasa<br>Pengurusan<br>Kolej | ~Jawatankuasa<br>Pengurusan<br>Kolej                                | ~Jawatankuasa<br>Pengurusan<br>Kolej                                  | ~Jawatankuasa<br>Pengurusan<br>Kolej                            | ~Jawatankuasa<br>Pengurusan<br>Kolej                   | ~Jawatankuasa<br>Pengurusan<br>Kolej                               | ~Jawatankuasa<br>Pengurusan<br>Kolej             | ~Jawatankuasa<br>Pengurusan<br>Kolej      | ~Jawatankuasa<br>Pengurusan<br>Kolej            | ~Jawatankuasa<br>Pengurusan<br>Kolej      | ~Jawatankuasa<br>Pengurusan<br>Kolej       | ~Jawatankuasa<br>Pengurusan<br>Kolej |
| Friday    | ~DCS651<br>~DCS651                   | ~DCS651<br>~DCS651                   | ~DCS651,<br>~Assignment 1<br>DCS651  | ~DCS651                                                             | ~DCS651,<br>~Assignment 2<br>DCS651                                   | ~DCS651                                                         | ~DCS651,<br>~Assignment 3<br>DCS651, ~Quiz<br>1 DCS651 | ~DCS651                                                            | ~DCS651,<br>~Assignment 4<br>DCS651              | ~DCS651,<br>~Test 1<br>DCS651             | ~DCS651,<br>~Assignment 5<br>DCS651             | ~DCS651,<br>~Quiz 2<br>DCS651             | ~DCS651,<br>~Assignment 6<br>DCS651        | ~DCS651                              |
| Saturday  | ~TKC501                              | ~TKC501                              | ~TKC501                              | ~TKC501                                                             | ~TKC501                                                               | ~TKC501,<br>~Assignment 1<br>TKC501                             | ~TKC501                                                | ~TKC501,<br>~Quiz 1<br>TKC501                                      | ~TKC501,<br>~Assignment 2<br>TKC501              | ~TKC501                                   | ~TKC501,<br>~Quiz 2<br>TKC501                   | ~TKC501, ~Test<br>1 TKC501                | ~TKC501, ~Test<br>2 TKC501                 | ~TKC501                              |
| Sunday    | ~Futsal                              | ~Futsal                              | ~Futsal                              | ~Futsal                                                             | ~Futsal                                                               | ~Futsal                                                         | ~Futsal                                                | ~Futsal                                                            | ~Futsal                                          | ~Futsal                                   | ~Futsal                                         | ~Futsal                                   | ~Futsal                                    | ~Futsal                              |

Fig. 4. Semester Planner using Greedy Algorithm Optimization.

TABLE III. RELIABILITY TESTING BASED ON ACCURACY TEST BETWEEN BRUTE FORCE AND TIMETABLE GENERATOR

| No | Brute Force Result, $p$                                                                                             |                                                                                                                    |                                                                                                                               | Timetable Generator Result, $p$                                                                                               |
|----|---------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
|    | Focus scale, $f$                                                                                                    | Complete time, $t$                                                                                                 | Priority, $p = f/t$                                                                                                           |                                                                                                                               |
| 1  | [2, ENT600]<br>[1, ITS610]<br>[1, IPK501]<br>[1, CSP650]<br>[1, ISEC]                                               | [2, ENT600]<br>[2, ITS610]<br>[2, IPK501]<br>[2, CSP650]<br>[2, ISEC]                                              | [1, ENT600]<br>[0.5, ITS610]<br>[0.5, IPK501]<br>[0.5, CSP650]<br>[0.5, ISEC]                                                 | [1, ENT600]<br>[0.5, ITS610]<br>[0.5, IPK501]<br>[0.5, ISEC]<br>[0.5, SCP650]                                                 |
| 2  | [6, ENT600]<br>[3, CSC662]<br>[3, ICT662]<br>[2, ITS610]                                                            | [5, CSC662]<br>[3, ENT600]<br>[2, ITS610]<br>[2, ICT662]                                                           | [1.5, ICT662]<br>[1.333, ENT600]<br>[1, ITS610]<br>[0.6, CSC662]                                                              | [1.5, ICT662]<br>[1.333, ENT600]<br>[1, ITS610]<br>[0.6, CSC662]                                                              |
| 3  | [10, Cooking]<br>[8, CSP650]<br>[7, ENT600]<br>[6, Brass Band]<br>[4, IPK501]<br>[3, ITS610]                        | [2, CSP650]<br>[2, ENT600]<br>[2, ITS610]<br>[2, IPK501]<br>[2, Brass Band]<br>[1, Cooking]                        | [10, Cooking]<br>[4, CSP650]<br>[3.5, ENT600]<br>[3, Brass Band]<br>[2, IPK501]<br>[1.5, ITS610]                              | [10, Cooking]<br>[4, CSP650]<br>[3.5, ENT600]<br>[3, Brass Band]<br>[2, IPK501]<br>[1.5, ITS610]                              |
| 4  | [10, CSP650]<br>[9, CSC445]<br>[8, ICT662]<br>[8, CSC548]<br>[6, Multimedia Club]<br>[6, Debates]<br>[2, Badminton] | [8, Badminton]<br>[6, CSP650]<br>[3, ICT662]<br>[2, Multimedia Club]<br>[2, Debates]<br>[1, CSC445]<br>[1, CSC548] | [9, CSC445]<br>[8, CSC548]<br>[3, Multimedia Club]<br>[3, Debates]<br>[2.667, ICT662]<br>[1.667, CSP650]<br>[0.25, Badminton] | [9, CSC445]<br>[8, CSC548]<br>[3, Multimedia Club]<br>[3, Debates]<br>[2.667, ICT662]<br>[1.667, CSP650]<br>[0.25, Badminton] |
| 5  | [10, ISP610]<br>[7, ACIS Club]<br>[6, CSC662]<br>[6, FSKM Club]<br>[4, DCS651]<br>[3, Reading]                      | [3, DCS651]<br>[3, ISP610]<br>[3, FSKM Club]<br>[2, Reading]<br>[1, ICT662]<br>[1, CSC662]                         | [7, ACIS Club]<br>[6, CSC662]<br>[3.333, ISP610]<br>[2, ICT662]<br>[2, FSKM Club]<br>[1.5 Reading]                            | [7, ACIS Club]<br>[6, CSC662]<br>[3.333, ISP610]<br>[2, ICT662]<br>[2, FSKM Club]<br>[1.5 Reading]                            |

|    |                                                                                                                                                   |                                                                                                                                                   |                                                                                                                                                                   |                                                                                                                                                                   |
|----|---------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|    | [2, ICT662]                                                                                                                                       | [1, ACIS Club]                                                                                                                                    | [1.333, DCS651]                                                                                                                                                   | [1.333, DCS651]                                                                                                                                                   |
| 6  | [10, CSP650]<br>[4, Watching movie]<br>[4, Reading]<br>[3, IPK501]<br>[3, ITS610]<br>[3, ENT600]<br>[2, ISEC]                                     | [4, Watching movie]<br>[2, IPK501]<br>[2, ITS610]<br>[2, CSP650]<br>[2, ENT600]<br>[2, Reading]<br>[1, ISEC]                                      | [5, CSP650]<br>[2, ISEC]<br>[2, Reading]<br>[1.5, IPK501]<br>[1.5, ITS610]<br>[1.5, ENT600]<br>[1, Watching movie]                                                | [5, CSP650]<br>[2, ISEC]<br>[2, Reading]<br>[1.5, IPK501]<br>[1.5, ITS610]<br>[1.5, ENT600]<br>[1, Watching movie]                                                |
| 7  | [10, CSP600]<br>[7, IPK501]<br>[7, Edit photo]<br>[6, JPNR]<br>[4, ENT600]<br>[4, ITS610]<br>[1, EET669]                                          | [8, ITS610]<br>[6, IPK501]<br>[6, CSP600]<br>[5, ENT600]<br>[3, JPNR]<br>[3, Edit photo]<br>[2, EET669]                                           | [2.333, Edit photo]<br>[2, JPNR]<br>[1.667, CSP600]<br>[1.167, IPK501]<br>[0.8, ENT600]<br>[0.5, ITS610]<br>[0.5, EET669]                                         | [2.333, Edit photo]<br>[2, JPNR]<br>[1.667, CSP600]<br>[1.167, IPK501]<br>[0.8, ENT600]<br>[0.5, ITS610]<br>[0.5, EET669]                                         |
| 8  | [6, CSP650]<br>[2, ENT600]<br>[2, ITS610]<br>[2, Studying]<br>[1, EET669]<br>[1, IPK501]<br>[1, ISEC]                                             | [8, Studying]<br>[3, CSP650]<br>[2, ITS610]<br>[2, IPK501]<br>[2, ISEC]<br>[1, EET669]<br>[1, ENT600]                                             | [2, ENT600]<br>[2, CSP650]<br>[1, EET699]<br>[1, ITS610]<br>[0.5, IPK501]<br>[0.5, ISEC]<br>[0.25, Studying]                                                      | [2, ENT600]<br>[2, CSP650]<br>[1, EET699]<br>[1, ITS610]<br>[0.5, IPK501]<br>[0.5, ISEC]<br>[0.25, Studying]                                                      |
| 9  | [10, Cooking]<br>[8, IPK501]<br>[8, CSP650]<br>[7, ENT600]<br>[6, Brass band]<br>[6, Watching movie]<br>[3, ITS610]<br>[3, Ping pong]             | [4, Watching movie]<br>[3, Ping pong]<br>[2, IPK501]<br>[2, CSP650]<br>[2, ENT600]<br>[2, ITS610]<br>[2, Brass band]<br>[1, Cooking]              | [10, Cooking]<br>[4, IPK501]<br>[4, CSP650]<br>[3.5, ENT600]<br>[3, Brass band]<br>[1.5, ITS610]<br>[1.5, Watching movie]<br>[1, Ping pong]                       | [10, Cooking]<br>[4, IPK501]<br>[4, CSP650]<br>[3.5, ENT600]<br>[3, Brass band]<br>[1.5, ITS610]<br>[1.5, Watching movie]<br>[1, Ping pong]                       |
| 10 | [9, DCS651]<br>[9, CSC649]<br>[9, CSC580]<br>[8, ICT663]<br>[4, JPK]<br>[3, TKC501]<br>[3, Futsal]<br>[2, CSC253]<br>[2, Netball]<br>[2, Fishing] | [5, DCS651]<br>[3, Netball]<br>[3, Fishing]<br>[2, TCK501]<br>[2, CSC253]<br>[2, ICT663]<br>[2, CSC649]<br>[2, CSC580]<br>[2, JPK]<br>[2, Futsal] | [4.5, CSC580]<br>[4.5, CSC649]<br>[4, ICT663]<br>[2, Futsal]<br>[2, JPK]<br>[1.8, DCS651]<br>[1.5, TKC501]<br>[1, CSC253]<br>[0.667, Netball]<br>[0.667, Fishing] | [4.5, CSC649]<br>[4.5, CSC580]<br>[4, ICT663]<br>[2, JPK]<br>[2, Futsal]<br>[1.8, DCS651]<br>[1.5, TKC501]<br>[1, CSC253]<br>[0.667, Netball]<br>[0.667, Fishing] |

TABLE IV. SUMMARY OF RESEARCH ACCURACY RESULT

| Result \ Test       | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|---------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Brute force         | 5   | 4   | 6   | 7   | 7   | 7   | 7   | 7   | 8   | 10  |
| Timetable Generator | 3   | 4   | 6   | 7   | 7   | 7   | 7   | 7   | 8   | 10  |
| Accuracy            | 0.6 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

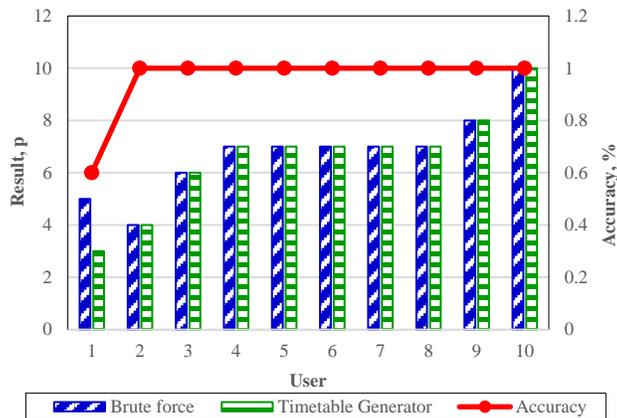


Fig. 5. Phases in Optimizing Greedy Algorithm for Timetable Generator.

## V. CONCLUSION

We presented a recommended semester planner using the optimization technique greedy optimization. Optimization is done to suggest a semester planner according to the user's preferences and at the same time recommend the planner with scheduling all user's activities, including study, hobby and extracurricular for a semester. We managed to validate the Timetable Generator using a reliability test by testing the system's accuracy with the Brute Force algorithm with the achievement of 100% accuracy. Future work is recommended to enhance the system's security. We need to hold the private and confidential data and propose using progressive web applications for a better view and suitable for handphones.

## ACKNOWLEDGMENT

The research was sponsored by Universiti Teknologi MARA Cawangan Melaka under the TEJA Grant 2021 (GDT 2021/1-28).

## REFERENCES

- [1] S. Singh and K. Kant, "Schools, skills, and learning: The impact of COVID-19 on education," 2020. <https://voxeu.org/article/impact-covid-19-education> (accessed Oct. 04, 2021).
- [2] N. B. Reinke, "The impact of timetable changes on student achievement and learning experiences," *Nurse Education Today*, vol. 62, pp. 137–142, 2018.
- [3] M. Huebener, S. Kuger, and J. Marcus, "Increased instruction hours and the widening gap in student performance," *Labour Economics*, vol. 47, pp. 15–34, 2017.
- [4] L. Foulkes, D. McMillan, and A. Gregory, "A bad night's sleep on campus: an interview to achieve exam success," *Sleep Health*, vol. 5, no. 3, pp. 280–287, 2019.
- [5] S. Panda, M. Mandal, and R. Barman, "Predictors of perceived stress among university students," *International Journal of Educational and Management Studies*, vol. 5, no. 4, pp. 324–328, 2015.
- [6] C. Kridiotis and S. Swart, "A learning development module to support academically unsuccessful 1st-year medical students," *African Journal of Health Professions Education*, vol. 9, no. 2, pp. 62–66, 2017.
- [7] A. M. Hays and S. Sharp, "Supporting postgraduate coursework students through their time of transition," *Research Online*, vol. 5, no. 2, pp. 40–54, 2018.
- [8] X. Brioso, A. Humero, and S. Calampa, "Comparing point-to-point precedence relations and location-based management system in last planner system: A housing project of highly repetitive processes case study," *Procedia Engineering*, vol. 164, pp. 12–19, 2016.
- [9] J. A. Grissom, S. Loeb, and H. Mitani, "Principle time management skills: Explaining patterns in principals' time use, job stress and perceived effectiveness," *Journal of Educational Administration*, vol. 53, no. 6, pp. 773–793, 2015.
- [10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. The Knuth-Morris-Pratt Algorithm, 2001.
- [11] S. Singh and K. Kant, "Greedy grid scheduling algorithm in dynamic job submission environment," in *International Conference on Emerging Trends in Electrical and Computer Technology*, 2011, pp. 933–936.
- [12] M. Choudhary and S. K. Peddoju, "A dynamic optimization algorithm for task scheduling in cloud environment," *International Journal of Engineering Research and Applications*, vol. 2, no. 3, pp. 2564–2568, 2012.
- [13] K. Wang, W. Shang, M. Liu, W. Lin, and H. Fu, "A greedy and genetic fusion algorithm for solving course timetable problem," in *17th International Conference on Computer and Information Science (ICIS)*, 2018, pp. 344–349.
- [14] S. Kadry and B. Ghazal, "New proposed design to solve examination timetable problem," in *3rd International Proceedings on Modern Trends in Science, Engineering and Technology*, 2015, pp. 36–40.
- [15] W. Shaochang, X. Fei, Y. Weixia, and M. Zhe, "Application of greedy random adaptive search algorithm (GRASP) in flight recovery problem," in *Second International Conference of Sensor Network and Computer Engineering (ICSNCE 2018)*, 2018, pp. 78–83.
- [16] Y. Tian, X. Zhang, C. Wang, and Y. Jin, "An evolutionary algorithm for large-scale sparse multiobjective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 380–393, 2019.
- [17] V. Kachitvichyanuku, "Comparison of three evolutionary algorithms: GA, PSO, and DE," *Industrial Engineering and Management Systems*, vol. 11, no. 3, pp. 215–223, 2012.
- [18] G. Prabhakaran, B. S. H. Khan, and L. Rakesh, "Implementation of grasp in flow shop scheduling," *The International Journal of Advanced Manufacturing Technology*, vol. 30, no. 11, pp. 1126–1131, 2006.
- [19] C. Cerrone, R. Cerulli, and B. Golden, "Carousel greedy: A generalized greedy algorithm with applications in optimization," *Computers & Operations Research*, vol. 85, pp. 97–112, 2017.
- [20] A. Levitin, *Introduction to the design & analysis of algorithm 3rd edition*. Pearson, 2012.
- [21] M. Mahoor, F. R. Salmasi, and T. A. Najafabadi, "A hierarchical smart street lighting system with brute-force energy optimization," *IEEE Sensors Journal*, vol. 17, no. 9, pp. 2871–2879, 2017.

# Development of an Efficient Electricity Consumption Prediction Model using Machine Learning Techniques

Ghaidaa Hamad Alraddadi<sup>1</sup>

Department of Computer Science  
College of Computer, Qassim University  
Buraydah 51452, Saudi Arabia

Mohamed Tahar Ben Othman<sup>2</sup>

BIND Research Group, IEEE Senior Member  
Department of Computer Science, College of Computer  
Qassim University, Buraydah 51452, Saudi Arabia

**Abstract**—Electricity consumption has continued to go up rapidly to follow the rapid growth of the economy. Therefore, detecting anomalies in buildings' energy data is considered one of the most essential techniques to detect anomalous events in buildings. This paper aims to optimize the electricity consumption in households by forecasting the consumption of these households and, consequently, identifying the anomalies. Further, as the used dataset is huge and published publicly, many research used part of it based on their needs. In this paper, the dataset is grouped as daily consumption and monthly consumption to compare the network topologies of all other works that used the same dataset with the selected part. The proposed methodology will depend basically on long short-term memory (LSTM) because it is powerful, flexible, and can deal with complex multi-dimensional time-series data. The results of the model can accurately predict the future consumption of individual households in a daily or monthly consumption base, even if the household was not included in the original training set. The proposed daily model achieves Root Mean Square Error (RMSE) value of 0.362 and mean absolute error (MAE) of 19.7%, while the monthly model achieves an RMSE value of 0.376 and MAE of 17.8%. Our model got the lowest accuracy result when compared with other compared network topologies. The lowest RMSE achieved from other topologies is 0.37 and the lowest MAE is 18% where our model achieved RMSE of 0.362 and MAE of 17.8%. Further, the model can detect the anomalies efficiently in both daily electricity consumption data and monthly electricity consumption data. However, the daily electricity consumption readings are way better to detect anomalies than the monthly electricity consumption readings because of the different picks that appear in the daily consumption data.

**Keywords**—Anomalies detection; deep learning; electricity consumption forecasting; LSTM

## I. INTRODUCTION

Global electricity consumption has grown rapidly faster than the rate of energy consumption where the electricity consumption in both commercial and residential buildings has significantly increased and can account between 20% and 40% in developed countries [1-2]. During 1980-2013, energy consumption went up from 7300 TWh to 22100 TWh. Also, it has grown even faster since the twenty-first century by 1.2 percentage points more than the average annual rise in energy consumption. In 2013, the annual electricity consumption of the world reached 3048 KWh per capita which is up to 42.3%

from 1990. In Asia, Bahrain, South Korea, and United Arab Emirates are the top three consumers where they exceeded 10000 KWh [2].

As a result, the energy demand will steadily be increased shortly due to the rise in population, comfort levels of buildings and spending a long time inside buildings. Thus, optimizing energy consumption and the efficiency of energy in buildings is a primary concern for anyone wishing to save energy [1]. Although the extensive modeling techniques that investigate designing buildings with a low level of energy consumption, buildings (especially commercial) often exceed the promised energy-saving design by some anomalous events, such as lighting equipment faults. These anomalous events can reach figures between 2-11% of the total energy consumption in commercial buildings [3].

To this end, detecting anomalies in buildings' energy data is considered one of the most essential techniques to detect anomalous events in buildings. Therefore, metering buildings' electricity provides data that have a significant impact on energy-saving opportunities due to analyzing energy usage, managing energy consumption and, thus, identifying anomalous patterns which if corrected will improve the comfort level of buildings with the least power consumption.

At a household level, the electricity consumption includes a high amount of noise, and the time series data is considered nonlinearly because of the seasonal changes' effects. Thus, this is a motivative of using machine learning field because of its capability to capture the nonlinearities among the time series data. In particular, LSTM models have proven effective in this type of context and in learning complex nonlinear patterns [4]. In this study, the aim is to solve the high volatility and uncertainty of electricity consumption in households by forecasting the consumption of these households and, consequently, identifying the anomalies.

Therefore, the proposed approach can deal effectively with a large number of smart meters data. After training and implementing the appropriate parameters, the resulting model will have the ability to forecast the future consumption of households that were not included in the training process. Hence, the resulting model has a great potential to forecast big data accurately. The validity of the proposed approach is tested based on an extensive real-world dataset that contains

thousands of households' consumption in several years so, several seasonal patterns. The used dataset is originally published in [5]. It is a collection of 104057 records of London's households from 2011 to 2014. In addition to the date and time, it consists of a unique household identifier, Acorn group categories, which is a classification of neighborhoods. Furthermore, historical weather data for London area during the same dates of smart meters registers was merged with the original dataset, which can be found in [6]. The weather dataset consists of the date of the day, maximum and minimum temperature of that day, high and low apparent temperature, wind bearing, dewpoint, cloud cover, wind speed, pressure, visibility, humidity, UV index, and moon phase.

The paper is organized as follows. Section II outlines historical selected works performed within electricity consumption forecasting. Then, the model is proposed in Section III. In Section IV, the results of the electricity consumption forecasting with the proposed model are discussed and shown. Finally, Section V concludes the paper and Section IV shows the future work.

## II. RELATED WORK

Recently, many machine learning models have contributed very well to estimating energy consumption by prediction models. The following section mentions some of the recently published researches in the areas of predicting energy consumption by using different machine learning approaches.

In references [7-9], they reviewed the state of the art of machine learning models of all kinds of energy systems including demand prediction, cost prediction, energy consumption prediction, load forecasting, etc. They identified the highest popularity models in energy systems which are Multilayer Perceptron (MLP), Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference System (ANFIS), Support Vector Machines (SVM), Extreme Learning Machine (ELM), Wavelet Neural Network (WNN), ensembles, deep learning, and hybrid machine learning models. Nevertheless, the hybrid machine learning models have significantly increased the performance of energy models.

Besides, García-Martín et al. illustrated in [10] the recent approaches used to estimate the energy consumption especially from data mining and neural networks perspectives. They revealed some emerged works, such as NeuralPower and SyNERGY that allow energy evaluation in machine learning. NeuralPower evaluates the energy by building a prediction model to estimate the energy consumption while SyNERGY builds energy estimation models based on integrating tools to the current machine learning suites. Further, they presented some challenges faced by current modeling approaches, such as the rapid changes in hardware, implementation and design of neural networks.

In [11] the authors proposed a hybrid approach that combines ELM with stacked autoencoders (SAE) to predict the energy consumption in buildings. Firstly, they used SAE to extract the features of energy consumption in buildings. Then, for prediction, they employed ELM to get precise prediction results. Their results showed that their proposed approach has

the best prediction performance compared to other approaches, such as SVM, multiple linear regression, backpropagation neural network (BPNN) and the generalized radial basis function neural network (GRBFNN). Also, in [12], they used deep extreme learning machine and SVM techniques to predict the energy consumption. Their proposed model obtained an accuracy of 90.70%.

J. Y. Kim and S. B. Cho. [13] proposed an autoencoder model to predict electricity consumption based on LSTM. The proposed model defines a state that represents the demand information, then, the future energy demand is predicted according to this state. The state contains information like the input values features, produced data features and the expected energy consumption. Moreover, this model allows inserting conditions to predict the electricity demand according to these conditions, such as economy or weather information and that is what makes it more efficient compared to other works. In addition to [13] and [11], Y. Jin *et al.* [14] proposed a clustering analysis method to analyze the daily electricity consumption based on an autoencoder algorithm. Their suggested method has a limitation of outlier detection when there are large outlier data.

In [15], they used different machine learning algorithms such as linear regression (LR), DTs, deep neural network (DNN), recurrent neural network (RNN), gated recurrent units (GRUs) and LSTM to evaluate their performance. They forecasted the data based on the ACORN groups in London. Then, they forecasted number of step-ahead like 1, 2, 12, 24, 48. LSTM achieved the lowest MAE compared to other algorithms for all forecasting types where they were 19%, 21.7%, 23.5%, 24.2%, and 25.6%, respectively. On the other hand, forecasting one step ahead has the lowest MAE and the worse was forecasting 48 steps ahead, as shown in Fig. 1. Furthermore, the authors in [16] proposed a machine learning-based ensemble model to improve the electricity consumption prediction. The model combines Cat Boost (CB), Gradient Boost (GB) and Multilayer Perceptron (MLP) algorithms. Moreover, they employed the genetic algorithm to get optimal features to be used for the model. They obtained RMSE of 5.05 and MAE of 3.05.

F. Z. Abera et al. proposed in [17] a method that uses the CLARA clustering technique to group their dataset into three clusters based on the mean of the consumption values. Then, SVM and ANN classifiers are used to predict the appliance that consumes more energy. They proved that ANN and SVM are worthwhile methods for analyzing and forecasting smart meter data with an accuracy of 99%.

| M  | Regression Model |       |       |       |       |              |
|----|------------------|-------|-------|-------|-------|--------------|
|    | LR               | DT    | DNN   | RNN   | GRU   | LSTM         |
| 1  | 26.2%            | 24.9% | 23.5% | 22.4% | 22.5% | <b>19.2%</b> |
| 2  | 27.6%            | 27.5% | 26.7% | 25.0% | 27.6% | <b>21.7%</b> |
| 12 | 30.2%            | 28.1% | 27.1% | 27.7% | 27.3% | <b>23.5%</b> |
| 24 | 30.0%            | 28.1% | 28.3% | 28.8% | 28.4% | <b>24.2%</b> |
| 48 | 30.7%            | 29.8% | 28.7% | 28.1% | 27.8% | <b>25.6%</b> |

Fig. 1. Forecasting Results of [15].

D. H. Nguyen et al. proposed in [18] a machine learning-based approach called iRBF-NN to predict the electricity consumption based on historical consumption and weather data. They demonstrated the relation between weather parameters (temperature, humidity, precipitation, sunshine duration and wind speed) and electricity consumption. They found that the humidity and temperature parameters have the highest relation to electricity consumption. Other parameters such as population and sunshine also affect the electricity consumption, however, according to the hardness of collecting their data and the simplicity of the model they did not consider them. The prediction performance of the proposed approach was good; however, the weather prediction was not accurate. E. Y. Shchetinin. [19] proposed a method to estimate the electricity consumption in commercial and business buildings by using a gradient boosting algorithm (GBM). Their results showed that GBM has the ability to estimate the accuracy of energy consumption prediction more than other machine learning algorithms like random forest and regression. Conversely, X. M. Zhang et al. used in [20] support vector regression to predict residential electricity consumption (rather than commercial consumption) according to their daily consumption and hourly consumption. Their results showed that the MAPE error is 12.78 and 22.01 for daily prediction and hourly prediction, respectively, and that means predicting daily consumption is better than predicting hourly consumption since it mitigates the effect of the randomness of behaviors in hourly family members.

S. Aman et al. proposed in [21] a novel model namely REDUCE (Reduced Electricity Consumption Ensemble) that combines outputs from three base models. These three models are called pre-DR (demand response), in-DR and all-day consumption sequences. Their results showed that the model is strong for buildings that do not follow a strict schedule of electricity consumption and they do not have enough historical demand response data.

On the other hand, Eisses, J. used in [22] three different machine learning techniques to detect anomalies in electricity consumptions data which are k-nearest neighbors (KNN), SVM and ANN. They conclude that ANN has the best accuracy performance with an error of 14%. Furthermore, K. Hollingsworth et al. [23] proposed an application for detecting anomalies in energy. Their application is based on the combination of two types of machine learning algorithms: ARIMA and LSTM. Their application correctly identified the anomalies and provided the time of the incident. Also, it provides higher accuracy by benefiting from both separated models' abilities. In [24], they combine K-means and DNN to identify the anomalies in energy consumption. Firstly, K-means is used to cluster the customers based on their similar electricity consumption behavior. Then, DNN algorithm is used to accurately identify the anomalies of each consumer.

### III. METHODOLOGY

Long short-term memory (LSTM) is a recurrent neural network (RNN) that is used widely in the deep learning field. RNN has the ability to process any hidden patterns that exist in the data since it takes into consideration the sequential nature

of the data. Therefore, it does not feed all the information to the network at once, but it feeds them as a chain structure where one element is processed and then passed to the second element in the sequence. In other words, RNN is recurrent in nature since it implements the same function for each input of the data whilst the output of the current input relies on the previous computation. Once the output is produced, it is copied and sent back into the recurrent network [25-26]. Fig. 2 shows the unrolled recurrent neural network where  $X_t$  is the input of the state,  $h_t$  is the output of the state and  $A$  is the activation function of the state. Firstly,  $X_0$  is taken from the sequence of the input and then outputs  $h_0$ . After that,  $h_0$  and  $X_1$  will be input for the next step. Similarly,  $h_1$  from the next step with  $X_2$  will be the input for the next step, etc.

Despite the stability of RNN, it has challenges in practice. It suffers from a well-documented problem called vanishing gradients.

The vanishing gradient is encountered when training a large number of samples since it requires many layers. Therefore, the gradient reduced dramatically because it propagated through the network [27].

LSTM can solve the issue of the vanishing gradient by capturing long-term dependencies. It's a well-known branch of deep learning and gained wide attention to forecasting time series data in recent years.

In this study, the aim is to solve the high volatility and uncertainty of electricity consumption in households by forecasting the consumption of these households and, consequently, identifying the anomalies.

#### A. Data Description and Preprocessing

The used dataset is real data of electricity consumption that was originally published in [5]. In particular, the dataset consists of electricity consumption readings of 5567 London households from November 2011 to February 2014. Readings measured in kWh were taken in half-hourly intervals. The dataset contains date and time, unique household identifier, Acorn group categories, which is a classification of neighborhoods. Furthermore, historical weather data for London area during the same dates of smart meters registers was merged with the original dataset, which can be found in [6]. The weather dataset consists of the date of the day, maximum and minimum temperature of that day, high and low apparent temperature, wind bearing, dewpoint, cloud cover, wind speed, pressure, visibility, humidity, UV index, and moon phase. Since the length of the samples varies from one household to another and it might that one household may have only one reading sample, we must assure that each household contains at least two readings one for each year, therefore, the others are excluded.

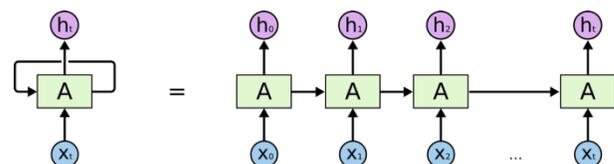


Fig. 2. Recurrent Neural Network.

As clarified earlier, the original electricity consumption data contains 167,932,474 lines of half-hourly readings for 5567 households in, approximately, four years. Nevertheless, there exist some missing consumption values. Therefore, to avoid the undesirable impact on the forecasting model data cleansing process should be done. Firstly, all NULL values are replaced with NAN. Meanwhile, there are 27858 of 0s in the consumption column where we consider them as inconsistent values. Therefore, they are replaced with NAN. After that, all NAN values are imputed by the average of the electricity consumption column. Fortunately, there are no missing or inconsistent values in the weather dataset. From the resulting dataset, 80% of the data are considered as training set and the other 20% for the testing set.

The proposed methodology is general and can deal with high-frequency time series data. However, the dataset is huge, and it contains more than 167 million records even after cleaning. Therefore, to reduce the dimension and volatility, the dataset is aggregated into daily intervals by calculating the mean consumption of each day for each household, as shown in Fig. 3. Moreover, for a comparison purpose, the average monthly electricity consumption is calculated for each household, as shown in Fig. 4. Fig. 3 illustrates the daily electricity consumption sorted by households' id. Although most of the consumption runs on average, there are many picks on different days that have high consumption. On the contrary, there is a clear difference between the monthly consumption in Fig. 4 and the daily consumption in Fig. 3 because when considering monthly consumption, it is mostly going to the average except for a huge pick in the middle. This potentially signs to get better results when using daily measurements for detecting anomalies rather than the monthly consumption because of the detailed data that appear in the daily readings.

### B. Proposed Model

Generally, the proposed methodology depends basically on LSTM because, as mentioned earlier, it is powerful, flexible, and has the ability to deal with complex multi-dimensional time-series data.

Further, an important reason behind using LSTM is that its performance is affected by the size of the data. Therefore, the main idea of the proposed model is to train a single model for all the considered data. The model is trained using long history to have the ability to predict new smart meters that were not used in the training process.

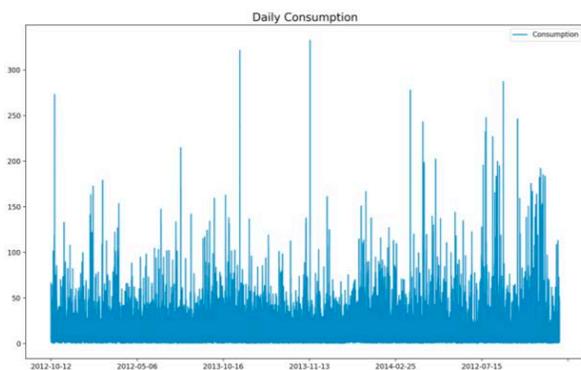


Fig. 3. Average Daily Electricity Consumption of London Households.

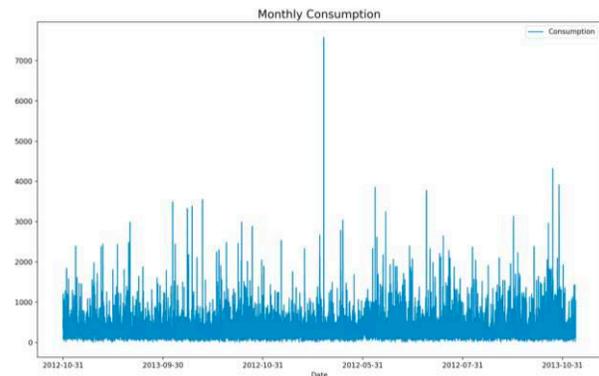


Fig. 4. Average Monthly Electricity Consumption of London Households.

To perform the electricity consumption prediction, firstly we set LSTM to do the time series forecasting with the historical data. Then, for detecting anomalies, the difference between the actual value and the predicted value will be calculated. A consumption is classified as an anomaly if its error is above a selected threshold. The threshold is selected intuitively depending on the density of the errors, as it will be clarified later.

To implement LSTM, the TensorFlow tool is used which is an open-source library developed to handle large datasets, optimization algorithms, and automatic differentiation. First, since the used data is time-series data, 3-dimensional units are defined to be used in the LSTM. The first dimension is the number of samples which is, in this case, the number of daily reading samples of the training set. The second dimension is the time steps that are used to forecast the current day (t) by given historical consumption and weather information at the prior time step, 30 in daily electricity consumption case and 15 in monthly electricity consumption case. The third and last dimension is the features which are the meteorological variables that it is defined earlier (like the weather conditions).

Next, the parameters of the model are adjusted as the following, the model is built with two LSTM layers. The first layer has 100 neurons and the second layer has 32 neurons. The 'relu' activation function is used to convert the output of each unit to be an input for the next layer. Further, a dropout layer is added with a rate of 20% to avoid overfitting. Finally, a dense layer is added with one unit to provide the corresponding forecast.

Immediately by designing the topology of the network, it compiles by defining the optimization algorithm and the loss function. the mean absolute value is selected for the loss function and Adam optimizer as the optimization algorithm. Indeed, all the previous parameters have been chosen after testing multiple network topologies. Once the network has been compiled, the associated weights are fitted which is a very expensive step in the methodology from a computational point of view. Then, the trained model is used to predict one step ahead for all the desired smart meters in the future.

The model needs to identify the number of epochs that indicate the number of passes of the entire training dataset. We select 40 epochs. Every epoch will be divided into a fixed-sized number from the training set, namely batch. In the daily

data, a batch size of 5000 days is selected since it is a big dataset. Hence, every epoch has 425 batches. However, for the monthly data, a batch size of 1000 days is selected, therefore, every epoch has 74 batches.

#### IV. RESULTS AND DISCUSSION

In this section, a summary of the main numerical results obtained from our proposed model in both daily electricity consumption and monthly electricity consumption is provided.

##### A. Electricity Consumption Forecasting

The forecasting model is designed to work with LSTM on both daily and monthly electricity consumption for each household. To explore the forecasting accuracy, two evaluation metrics are used are the root mean squared error (RMSE) and the mean absolute error (MAE) which can be calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y - \hat{Y})^2} \quad (1)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |Y - \hat{Y}| \quad (2)$$

In general, the smaller values of the performance metrics provide higher forecasting accuracy. Table I presents the forecasting results obtained from the proposed model and compares it with other works presented in [28], [15], [29] and [30], where the same dataset is shared. However, the used dataset is a huge dataset and its size is about 11 GB. Therefore, every research used part of it based on their needs, as illustrated in the second column of Table I. In this work, daily electricity consumption and monthly electricity consumption datasets are used to explore how the performance will be affected when using different network topologies in the same dataset. As it is clear from Table I, the proposed dataset and the dataset used in [28] are the most complicated and longest since we have not excluded any reading from the original dataset. However, in [28], they used only three weather conditions which are humidity, wind speed, and temperature whereas, in this work, all the weather conditions are used without excluding any influencing factor. The lowest RMSE value is achieved from [28] where it is equal to 0.05. On the other hand, the worst RMSE value obtained from [29] where is 3.35. To the research that used MAE performance measure, the lowest MAE was achieved by [16] when forecasting one step-ahead where they got MAE of 19%, whereas the worst MAE was obtained from the same research when forecasting 48 steps-ahead where they got MAE of 25.6%. Although the proposed model does not provide the best result when compared with other works, it got the lowest accuracy result when compared with other compared network topologies, as shown in the 12 and 13 columns in Table I.

To give a graphical representation of the daily electricity consumption model's results, Fig. 5 illustrates the forecasted consumption in comparison with the original consumption in both training and test data. As it is clear from the figure, there is some difference between the original values and the predicted ones. This might be because of the different

consumption habits of every household which increase and decrease without a clear pattern, as illustrated earlier in Fig. 3.

Though the proposed model appears as it did not forecast the daily electricity consumption perfectly, Fig. 6 can approve that the electricity consumption pattern forecasted well. In Fig. 6, the plot of test loss drops below training loss which means the model overcomes the overfitting problem and learned perfectly. Moreover, the proposed model achieves RMSE value of 0.362 and MAE of 19.7% which are the lowest values achieved when compared with other network topologies used in Table I.

After implementing the forecasting model for the monthly electricity consumption dataset, Fig. 7 shows the graphical representation of the true monthly electricity consumption and the forecasted values for both the training and testing process.

As it is clear, the graphical representation forecasting result of the monthly electricity consumption is better than the graphical representation of the daily electricity consumption prediction in Fig. 5. This approves the assumption that the daily electricity consumption contains a lot of high and low picks in the consumption which is contrary to the monthly electricity consumption data that run mostly in the average pattern.

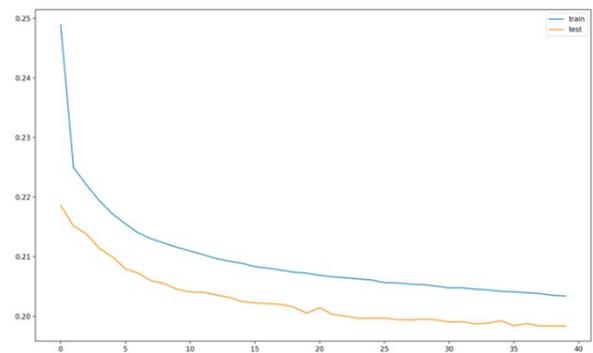


Fig. 5. Original Daily Electricity Consumption Data and Prediction Results of LSTM.

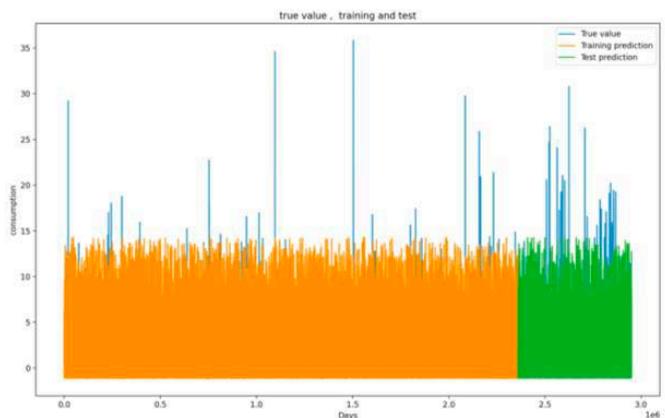


Fig. 6. The Training and Testing Loss of Daily Electricity Consumption Prediction for 40 Epochs.

TABLE I. A COMPARISON OF THE FORECASTING RESULTS OF DIFFERENT NETWORK TOPOLOGIES

| Paper          | Used data                                                               | Batch size | number of epochs | Optimizer | LSTM layers | LSTM units for each layer                                      | Activation function | Loss function | Forecasting type                                       | Performance               |       | Proposed daily dataset performance | Proposed monthly dataset performance |    |       |
|----------------|-------------------------------------------------------------------------|------------|------------------|-----------|-------------|----------------------------------------------------------------|---------------------|---------------|--------------------------------------------------------|---------------------------|-------|------------------------------------|--------------------------------------|----|-------|
|                |                                                                         |            |                  |           |             |                                                                |                     |               |                                                        | m                         | MAE   |                                    |                                      | m  | MAE   |
| [28]           | All buildings with parameters of humidity, wind speed, and temperature. | -          | -                | Adam      | 2           | 32 for each layer                                              | -                   | MAE           | Forecasting one week ahead                             | RMSE: 0.050               |       | RMSE: 0.492                        | RMSE: 0.37                           |    |       |
| [15]           | 50 households from 16 Acorn of the year of 2013                         | -          | 100              | -         | 3           | 32 for each layer                                              | tanh                | -             | Forecasting m steps ahead where m = [1, 2, 12, 24, 48] | m                         | MAE   | m                                  | MAE                                  | m  | MAE   |
|                |                                                                         |            |                  |           |             |                                                                |                     |               |                                                        | 1                         | 19.2% | 1                                  | 21%                                  | 1  | 18.2% |
|                |                                                                         |            |                  |           |             |                                                                |                     |               |                                                        | 2                         | 21.7% | 2                                  | 23.4%                                | 2  | 23.5% |
|                |                                                                         |            |                  |           |             |                                                                |                     |               |                                                        | 12                        | 23.5% | 12                                 | 27.3%                                | 12 | 46.3% |
|                |                                                                         |            |                  |           |             |                                                                |                     |               |                                                        | 24                        | 24.2% | 24                                 | 29.94%                               | 24 | 49%   |
| 48             | 25.6%                                                                   | 48         | 33.94%           | 48        | 46.6%       |                                                                |                     |               |                                                        |                           |       |                                    |                                      |    |       |
| [29]           | Half hourly consumption data of 500 days for 112 households             | 1          | 50               | -         | 4           | -                                                              | sigmoid             | MSE           | -                                                      | RMSE: 3.35                |       | RMSE: 0.45912                      | RMSE: 0.509                          |    |       |
| [30]           | Hourly consumption data for 3891 households of the year 2013            | 1000       | 40               | Adam      | 2           | 32 units for the first layer and 16 units for the second layer | tanh                | MAE           | Forecasting 24-hours ahead                             | MAE: 0.04                 |       | MAE: 0.206                         | MAE: 0.18                            |    |       |
| Proposed model | Daily electricity consumption data                                      | 5000       | 40               | Adam      | 2           | 100 for the first layer and 32 for the second layer            | ReLU                | MAE           | Forecasting one day ahead                              | MAE: 19.7%<br>RMSE: 0.362 |       | -                                  | -                                    |    |       |
| Proposed model | Monthly electricity consumption data                                    | 164        | 40               | Adam      | 2           | 100 for the first layer and 32 for the second layer            | ReLU                | MAE           | Forecasting one day ahead                              | MAE: 17.8%<br>RMSE: 0.376 |       | -                                  | -                                    |    |       |

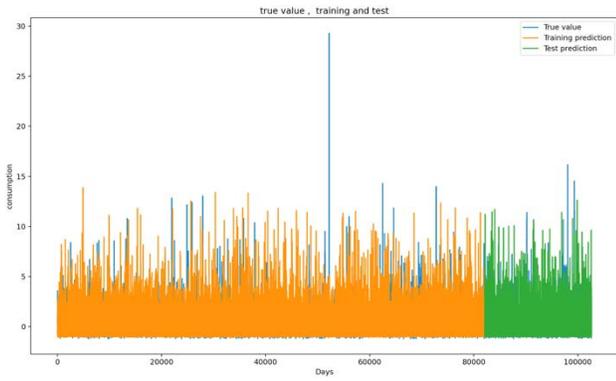


Fig. 7. Original Monthly Electricity Consumption Data and Prediction Results of LSTM.

Further, the loss plot of the daily electricity consumption data and the loss plot of the monthly electricity consumption data drop below training loss which means the model overcomes the overfitting problem and learned well for both datasets, as shown in Fig. 6 and Fig. 8.

Besides, the model of the monthly electricity consumption data achieves RMSE value of 0.376 and MAE of 17.8%. The RMSE value of the monthly electricity consumption is higher than the RMSE value achieved from the daily electricity consumption model. However, The MAE value for the monthly electricity consumption model is lower than the MAE of the daily electricity consumption model.

### B. Anomalies Detection

In the previous section, a model has been built to accurately forecast electricity consumption. Now, this model can be used to identify the anomalies in all considered data. The main idea is to forecast the electricity consumption at time  $t$ . Then, the difference between the actual value and the predicted value is calculated. A consumption is classified as an anomaly if its error is above a selected threshold. The threshold is selected intuitively depending on the density of the errors, as shown in Fig. 9 and Fig. 10. Fig. 9 illustrates the plots of the mean absolute error values of the daily electricity consumption data. Here, as shown in the figure the density of the errors is around zero and it is going up to 25. The large errors of the daily electricity consumption forecasting occur because of metering the different daily consumption habits of every household.

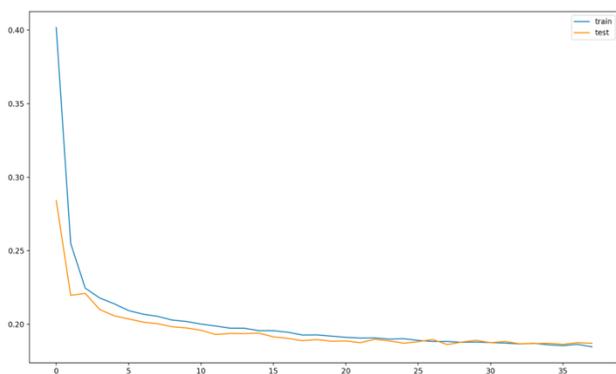


Fig. 8. Training and Testing Loss of Monthly Electricity Consumption for 40 Epochs.



Fig. 9. Mean Absolute Error between the Original Consumption's Value and the Predicted Value of the Daily Electricity Consumption Data.

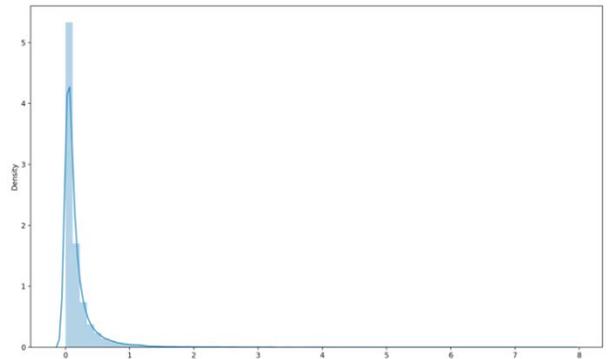


Fig. 10. Mean Absolute Error between the Original Consumption's Value and the Predicted Value of the Monthly Electricity Consumption Data.

Further, the density of the error of the monthly electricity consumption data is around zero; however, it is going up to 8 which is much lower than the errors in the daily consumption forecasting, as illustrated in Fig. 10.

After trying several attempts and adjustments, the proper threshold found is 7 for the daily electricity consumption data and 3 for the monthly electricity consumption data. Fig. 11 and Fig. 12 show the anomalies detected in the daily electricity consumption and the monthly electricity consumption, respectively. The blue line is the consumption, and the red dots are the anomalies.

Generally, the consumption is concerned as an anomaly because of the unexpected trend changes in the data like the sudden increase of the consumption that is different from the normal consumption.

As you can see from Fig. 11 and Fig. 12, it is clear that the model can detect the anomalies efficiently in both daily electricity consumption data and monthly electricity consumption data. However, the daily electricity consumption readings are way better to detect anomalies than the monthly electricity consumption readings because of the different picks that appear in the data. It has been already assumed that in Fig. 3 and Fig. 4 the daily readings will impact the forecasting results and that was true. The daily smart metering has larger errors between the true values and the predicted values than the monthly smart metering which helps detect the anomalies in the data.

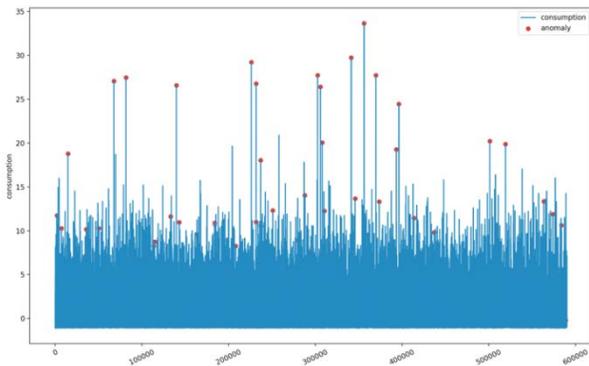


Fig. 11. Detected Anomalies in Daily Electricity Consumption Data.

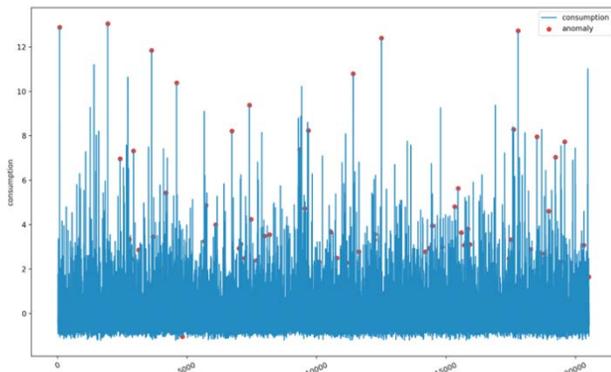


Fig. 12. Detected Anomalies in Monthly Electricity Consumption Data.

To conclude, both daily readings and monthly readings provide similar forecasting performance, however, using daily readings can provide detailed data for households more than the monthly readings. Therefore, using daily metering provide effective results for detecting the anomalies.

## V. CONCLUSION

This work is focusing on detecting anomalies in electricity consumption data by using the long short-term memory (LSTM) approach. The anomalies are identified in two steps: forecasting future consumption and thus anomalies detection. The proposed model is tested using a large real-world dataset with thousands of households segregated into daily consumption and monthly consumption to explore how these may impact the forecasting accuracy of the model. Since the used dataset is huge and published publicly, many research used part of it based on their needs. In this work, we did not exclude any information from the dataset. Instead, the average daily consumption and the average monthly consumption are calculated for comparison purposes.

In conclusion, the proposed model got the lowest accuracy result when compared with other network topologies. The lowest RMSE achieved from other topologies is 0.37 and the lowest MAE is 18% where the proposed model achieved RMSE of 0.362 and MAE of 17.8%. Moreover, both daily and monthly readings have similar forecasting performance; however, the daily readings provide more detailed data for households than the monthly readings. Therefore, using daily metering provides effective results to detect anomalies.

## VI. FUTURE WORK

In this work, the used dataset is public electricity consumption data. In the future, we aim to collect the electricity consumption data of Saudi Arabia's buildings. Furthermore, the weather information of the collected years will be added. Finally, there is a plan to construct an efficient energy management system that identifies the anomalies in daily real-time buildings' electricity consumption.

## ACKNOWLEDGMENT

The authors would like to thank Qassim University for supporting this research.

## REFERENCES

- [1] L. Pérez-Lombard, J. Ortiz and C. Pout, "A review on buildings energy consumption information," *Energy Build*, vol. 40, no. 3, pp. 394-398, 2008.
- [2] Z. Liu, "Global energy development: the reality and challenges," in *Global Energy Interconnection*, Academic Press, pp. 1-64, 2015.
- [3] Y. Heo, R. Choudhary and G. A. Augenbroe, "Calibration of building energy models for retrofit analysis under uncertainty," *Energy Build*, vol. 42, pp. 550-560, 2012.
- [4] K. Bandara, C. Bergmeir, and S. Smyl. "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," arXiv preprint arXiv:1710.03222, 2017.
- [5] [Online]. Available: <https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households>.
- [6] [Online]. Available: <https://www.kaggle.com/jeanmidev/smart-meters-in-london>.
- [7] A. Mosavi, M. Salimi, S. F. Ardabili, T. Rabczuk, S. Shamsirband and A. Varkonyi-Koczy, "State of the art of machine learning models in energy systems, a systematic review," *Energies*, vol. 12, no. 7, 2019.
- [8] M. Bourdeau, X. qiang Zhai, E. Nefzaoui, X. Guo and P. Chatellier, "Modeling and forecasting building energy consumption: A review of data-driven techniques," *Sustainable Cities and Society*, vol. 48, pp. 101-533, 2019.
- [9] U. Farouk, M. Asante and J. Ben, "A survey of machine learning's electricity consumption models," *Communications on Applied Electronics*, vol. 7, no. 21, pp. 6-10, 2018.
- [10] E. García-Martín, C. F. Rodrigues, G. Riley and H. Grahm, "Estimation of energy consumption in machine learning," *Journal of Parallel and Distributed Computing*, vol.134, pp. 75-88, 2019.
- [11] C. Li, Z. Ding, D. Zhao, J. Yi and G. Zhang, "Building energy consumption prediction: An extreme deep learning approach," *Energies*, vol. 10, no. 10, pp. 1-20, 2017.
- [12] T. M. Ghazal, S. Noreen, R. Said, M. Khan, S. Siddiqui, S. Abbas, S. Aftab and M. Ahmad, "Energy demand forecasting using fused machine learning approaches," *Intelligent Automation and Soft Computing*, vol. 31, no. 1, pp. 539-553, 2022.
- [13] J. Y. Kim and S. B. Cho, "Electric energy consumption prediction by deep learning with state explainable autoencoder," *Energies*, vol. 12, no. 4, 2019.
- [14] Y. Jin, D. Yan, X. Zhang, M. Han, X. Kang, J. An and H. Sun, "District household electricity consumption pattern analysis based on auto-encoder algorithm," *IOP Conference Series: Materials Science and Engineering*, vol. 609, no. 7, pp. 072-028, 2019.
- [15] P. A. Schirmer, I. Mporas and I. Potamitis, "Evaluation of regression algorithms in residential energy consumption prediction," in *3rd European Conference on Electrical Engineering and Computer Science, EECS 2019*. pp. 22-25, 2019.
- [16] P. W. Khan and Y. C. Byun, "Adaptive error curve learning ensemble model for improving energy consumption forecasting," *Computer, Material and Continua*, vol. 69, no. 2, pp. 1893-1913, 2021.
- [17] F. Z. Abera and V. Khedkar, "Machine learning approach electric appliance consumption and peak demand forecasting of residential

- customers using smart meter data," *Wireless Personal Communications*, vol. 111, no. 1, pp. 65–82, 2020.
- [18] A. H. Neto and F. A. S. Fiorelli, "Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption," *Energy Build.*, vol. 40, no. 12, pp. 2169–2176, 2008.
- [19] E. Y. Shchetinin, "Modeling the energy consumption of smart buildings using artificial intelligence," *CEUR Workshop Proc.*, vol. 2407, pp. 130–140, 2019.
- [20] X. M. Zhang, K. Grolinger and M. A. M. Capretz, "Forecasting residential energy consumption using support vector regressions," in *Proc. IEEE Inter. Conf. Mach. Learn. Appl.*, pp. 1–10, 2018.
- [21] S. Aman, C. Chelmis, and V. K. Prasanna, "Learning to REDUCE: A reduced electricity consumption prediction ensemble," *AAAI Work. - Technical Report*, vol. WS-16-01-, pp. 204–210, 2016.
- [22] J. Eisses, "Anomaly detection in electricity consumption data of buildings using predictive models," *University of Amsterdam*, pp. 1-7, 2014.
- [23] K. Hollingsworth, K. Rouse, J. Cho, A. Harris, M. Sartipi, S. Sozer, B. Enevoldson, "Energy anomaly detection with forecasting and deep learning," in *Proc. - 2018 IEEE Int. Conf. Big Data, Big Data* , pp. 4921–4925, 2019.
- [24] A. Maamar and K. Benahmed, "A hybrid model for anomalies detection in ami system combining k-means clustering and deep neural network," *Computer, Material and Continua*, vol. 60, no. 1, pp. 15–39, 2019.
- [25] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, pp. 132–306, 2020.
- [26] R. DiPietro and G. D. Hager, "Deep learning: RNNs and LSTM," *Handbook of Medical Image Computing and Computer Assisted Intervention*. pp. 503–519, 2019.
- [27] S. Hochreiterf, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107-116, 1998.
- [28] T. T. Q. Nguyen, T. P. T. Tran, V. Debusschere, C. Bobineau and R. Rigo-Mariani, "Comparing high accurate regression models for short-term load forecasting in smart buildings," in *Proc. (Industrial Electron. Conf.)*, pp. 1962–1967, 2020.
- [29] D. Kaur, R. Kumar, N. Kumar and M. Guizani, "Smart grid energy management using RNN-LSTM: A deep learning-based approach," in *proc. 2019 IEEE Global Communications Conference*, pp. 1-6, 2019.
- [30] A. M. Alonso, F. J. Nogales and C. Ruiz, "A single scalable lstm model for short-term forecasting of massive electricity time series," *Energies*, vol. 13, no. 20, 2020.

# Critical Review of Technology-Enhanced Learning using Automatic Content Analysis

## Case Study of TEL Maturity Assessment Formulation

Amalia Rahmah<sup>1</sup>, Harry B. Santoso<sup>2</sup>  
Faculty of Computer Science  
Universitas Indonesia  
Depok, Indonesia

Zainal A. Hasibuan<sup>3</sup>  
Faculty of Computer Science  
Universitas Dian Nuswantoro  
Semarang, Indonesia

**Abstract**—Technology-enhanced learning (TEL) continues to grow gradually while considering a multitude of factors, which underpins the need to develop a TEL maturity assessment as a guideline for this gradual improvement. This study investigates the potential application of TEL's expert knowledge presented in various research articles as qualitative data for developing assessment questionnaires. A mixed-method approach is applied to analyze the qualitative data using systematic literature review (SLR) with automated content analysis (ACA) as quantitative data processing to strengthen the trustworthiness of the findings and reduce researcher bias. This process is carried out six steps: conducting SLR, data processing with ACA using Leximancer, organizing resulting concepts with facet analysis, contextualizing each TEL facet, constructing the assessment questionnaire for each context, and establishing TEL maturity dimensions. This study generates 64 questionnaire statements grouped according to the target respondents, namely students, teachers, or institutions. This set of questions is also grouped into dimensions representing aligned context: student performance, learning process, applied technology, contents, accessibility, teachers and teachings, strategy and regulation. Further research is required to distribute this questionnaire for pilot respondents to design the improvement roadmap and check data patterns to formulate maturity appraisals and scoring methods.

**Keywords**—Automatic content analysis; ACA; assessment questionnaire; concept; facet analysis; key terms; Leximancer; systematic literature review; SLR; technology-enhanced learning; TEL; text analysis; theme

### I. INTRODUCTION

Technology-enhanced learning (TEL) exploits technological advancement for continuous learning improvement. However, these advances cannot always be applied simultaneously in every region or educational institution. Several factors influence TEL application success, including accessible technology, supporting infrastructure, the conditions of learners, teachers, and the institution where learning takes place. These factors imply that TEL must be applied at different rates but gradually improved. Therefore, guidance for the application of TEL is required, as conceptualized in the TEL maturity model with its assessment instruments.

This research generates an assessment questionnaire grouped in related dimensions to build this instrument, which

constructs a TEL maturity model. In related works, questionnaire formulations are derived from similar questionnaires in existing research, such as the capability maturity model (CMM), as seen in the maturity model for mobile learning [1]. Another example is a study about digital game maturity, which does not use CMM but still formulates the maturity's instrument based on the defined game development process [2] [3]. However, the maturity referred to in this study, namely the conditions for TEL application and how to gradually improve it, is inconsistent with maturity assessments emphasizing the maturity process. The other techniques in related work utilize qualitative analysis to exploit experts' knowledge or implement literature reviews from previous articles with a similar topic [4]. Combining these two techniques opens the prospect of developing TEL assessment instruments consistent with the previously determined TEL maturity context.

Regardless, issues with previous studies concern researcher bias affecting the process and the reliability of the terms used during the coding process in the qualitative analysis. However, in a deeper view, TEL practitioners and experts have made their knowledge explicit in various research journals. Thus, this study employs a literature review approach investigating experts' knowledge captured within related research articles.

The most common method for conducting a literature review is by searching for relevant articles. For example, this research has searched for various studies containing the keyword "TEL" and specifically discussing affecting factors, TEL assessment, or certain technology maturity. However, this process is considered insufficient because the obtained articles' scope of discussion does not meet the previously defined TEL maturity context. Furthermore, the existing literature review approach has shortcomings, such as the method's reliability in finding relevant articles, the possibility for researcher bias, and cognitive limitations in extracting knowledge from the vast amount of available research articles. Moreover, the requirement to assess TEL maturity requires understanding the entire scope of the TEL discussion. Thus, a systematic literature review (SLR) approach may reduce researcher bias and meet the requirement for a TEL maturity assessment.

In SLR-related works, statistical data processing has used only specific attributes from the selected articles. However, there is potential to uncover the underlying experts' and

researcher's knowledge in those articles. Additional approaches include research that seeks to automate the process of qualitative analysis [5]. Some of these approaches underlie why this study exploits automated content analysis (ACA) as a text analysis method using a tool called Leximancer. ACA produces a set of concepts (key terms) and themes (clustered concepts based on relevance). Previously, interpretation of ACA results, particularly those using Leximancer, have been limited to describing a topic's state of the art or research trends, as shown in [6], [7]. Nevertheless, the potential is enormous. Depending on how to organize the resulting knowledge, this collection of concepts can be interpreted from various perspectives.

Facet analysis is one method for organizing knowledge [8], including data that form a collection of concepts. As a cross-disciplinary approach, facet analysis is primarily used as a library classification method. Facet analysis can help represent the content of a broad discussion covering many documents. This study analyzes TEL facets, which can then be interpreted in various ways depending on the research objective, which in this work is to construct assessment instruments and dimensions for TEL maturity. Thus, this research examines the research question of how SLR and ACA can be used to develop a TEL assessment instrument.

The following five sections are structured to address the research question. Section 2 describes the rationale for this research's importance. Section 3 explains underlying theories and concepts to understand the research context. Then, Section 4 presents the methodology of conducting SLR using ACA, including structuring and interpreting the result. The following section discusses the interpretation of the previous step's results, which later become assessment questionnaires. The last section concludes the study and suggests further work to refine the instrument into a complete working framework for TEL assessment.

## II. RATIONALE

### A. Requirement for a TEL Maturity Assessment

Technology continues to advance, including learning technology. Various initiatives in implementing technological advances are also increasing, focusing on digitization, process acceleration, and learning improvement. The continuous application of technology advancement is the underpinning principle of technology-enhanced learning. According to Kirkwood [9], TEL represents the use of information and communication technology (ICT) for learning and teaching. Programs for implementing technology to improve learning typically use a top-down mechanism, which refers to how policymakers implement programs at the operational level. However, these programs do not always consider preexisting conditions such as the availability of supporting infrastructure and differences in students' abilities to access and use technology.

As a result, we require guidelines for implementing TEL that consider a variety of factors and gradual improvement. This gradual mechanism is encapsulated in the concept of a maturity model and its assessment instruments. However, Nicoll et al. [10] stated that research articles exploring TEL

evaluation remain limited. This study attempts to build this maturity model, which is intended for TEL evaluation and a gradual improvement guide.

The TEL maturity model has essential components including model domains, attributes, appraisal and scoring methods, and improvement roadmaps [11]. Two components are constructed: (1) an assessment questionnaire as model attributes and (2) dimensions as model domains. The concerning issue is establishing these attributes and domains considering that TEL covers an enormous scope of discussion. The subsequent issue is what point of view objectifies the TEL maturity concept. Thus, this study implements a qualitative analysis that explores the underlying knowledge from the entire scope of TEL discussion.

Previous research has attempted to formulate a conceptual framework for TEL, incorporating discussions about technology and learning [12]. However, there are two drawbacks: the object discussed is e-learning, where the term is not quite suitable for the research context. The second is that the result is insufficient to be further analyzed as a working framework. Thus, this study tries to answer these shortcomings.

Moreover, TEL is a broad field of study as it covers both technology and learning discussions. Therefore, it is difficult to determine the scope of discussion of TEL, particularly if the purpose of the research is to find the dimensions of TEL. This challenge suggests a systematic literature review, as explained in the next section.

### B. SLR with ACA to Gain the Whole Scope of TEL Discussion

The research context is the TEL domain, which includes discussions on technology's progression and application to learning advancement. The context is building a model that evaluates TEL maturity to recommend a strategy to improve the impact of technology use on learning. Then, the challenge is to find the factors representing TEL as a subject matter, which become the objects to be measured. Extracting insightful knowledge through literature reviews is an attempt to overcome this challenge.

Determining the factors to be measured is typically done by collecting knowledge from experts using the focus group discussion method or in-depth interviews. Meanwhile, every scientific publication article in renowned conferences and journals is also a form of knowledge externalization from experts or knowledgeable people in the TEL domain. This knowledge base justifies why the literature review may be used to acquire knowledge, followed by knowledge organization, with the result used to develop TEL maturity instruments.

According to Kitchenham [13], a systematic literature review is essential because evidence-based rather than expert opinion is also required. The article states that evidence can be in the form of a synthesis of best-quality scientific studies on a particular topic. In contrast to an expert review using ad hoc literature selection, an SLR is a methodologically rigorous review of research results. The objective of SLR is not merely to aggregate all existing evidence on a research question, it is also intended to support the development of evidence-based guidelines for practitioners.

The SLR uses specific criteria to select the papers to be reviewed. The review process is followed by statistical analysis and interpretation, resulting in understanding the subject under discussion. However, the literature review should draw on substantial knowledge suggested in the articles. Also, there is a requirement to extract knowledge in model development research that can be justified, reflecting the more thorough subject-matter discussion.

This research expands the technique of implementing quantitative data analysis on SLR by processing the entire content of the text from research articles rather than only particular attributes. An ACA approach is used to target this goal, employing the Leximancer tool. The concept of ACA is derived from text analysis, where all the terms in the various data source articles are parsed, cleaned, assigned weights, and sorted. Then, the relevance of each term to each other is calculated. Leximancer helps data processing visualize conceptual maps by generating main concepts contained within the text and determining how they are related. ACA with Leximancer is usually utilized for sentiment and data marketing analysis, as seen in [14] and [15]. Though, the resulting concepts and themes may deliver many useful insights with further analysis. It is then become one of the background idea in this research.

Regarding ACA, similar activities have been done in previous works using simpler text analysis techniques, as shown in Fig. 1. The first research utilized text analysis with the help of mini program built using Python, with a resulting list of key terms sorted by frequency of occurrence. Regarding interpretation of results, the previous study used Luhn's theory, which describes the relation of curves to Bradford-Zipf distributions to determine significant words [16]. These sorted words are located in between the upper and lower cut in the Zipf law distribution. In more recent research, the determination of the upper cut and lower cut was still not clearly stipulated, though a pattern was seen [17]. This provision of Luhn's theory was again adopted in this study. In the Methodology section, we discuss how this determination is established. The result of this study is a collection of TEL characteristics. However, this result cannot be directly converted into TEL assessment instruments.

Another work attempted to adopt the term frequency-inverse document frequency method [18] rather than a simple frequency count as a basis to sort the resulting list of key terms. However, this study also has a limitation in that it processed only 100 articles; therefore, its suitability for this research objective is dubious. Another study has already employed ACA [19], however, a drawback in the work is a lack of clarity on interpreting the results. In that study, the dimensions were directly derived from themes generated from Leximancer without further analysis, yet the underlying knowledge lies in the concepts generated. The resulting list of concepts requires different processing methods, which is what underlies the use of facet analysis.

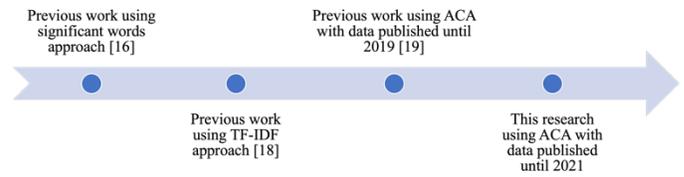


Fig. 1. This Research Series Timeline.

### C. Facet Analysis for Knowledge Interpretation and Organization

The output of ACA is a collection of key terms that are considered the most relevant and represent the scope of the TEL discussion, called the concept. However, these results require further analysis to establish dimensional candidates and assessment questionnaires. Quantitative data processing, automated with Leximancer, needs to be counteracted with interpreting the result, which is the significant proposal of this study.

Facet analysis departs from the classification theory originated by Ragnathan [20]. The technique recognizes several aspects of a topic of discussion and summarizes these aspects to appropriately describe the concept. A "facet" is a collection of terms that have the same relationship with the global subject, reflecting the application of a fundamental principle of division. Certain subjects are delivered in various perspectives so that the representation of knowledge can be in its entirety, not as a subordinate of something else.

According to Usman et al., facet analysis is one of the most frequently used classification structures in education and computer science [21]. In addition, the discussion of TEL is multidimensional and multiperspective (facets). Facet analysis is also widely studied in social science (21%), and computer science (11%), both of which are also domains that frequently examine TEL. Thus, facet analysis is not new in computer science and may be suitable to be applied in this study.

## III. METHODOLOGY

This study addresses the research question by following the steps depicted in Fig. 2. The diagram shows how each step has inputs and outputs and employs particular techniques. In the following subsections, each step is explained.

### A. SLR for Data Gathering

After various data selection processes applied to 1,030 articles filtered by inclusion, exclusion, and quality-assessment criteria, 792 journal articles were obtained. Inclusion and exclusion criteria include Scopus-indexed journals, in English, published between 2010 and 2021, available for full-text download, and mentioning "technology-enhanced learning" in the titles or abstracts. The process continued by performing text analysis on the entire content of the selected articles using the ACA method. The aim of the analysis was to explore underlying knowledge representing the overall scope of the TEL discussion rather than only specific attributes.

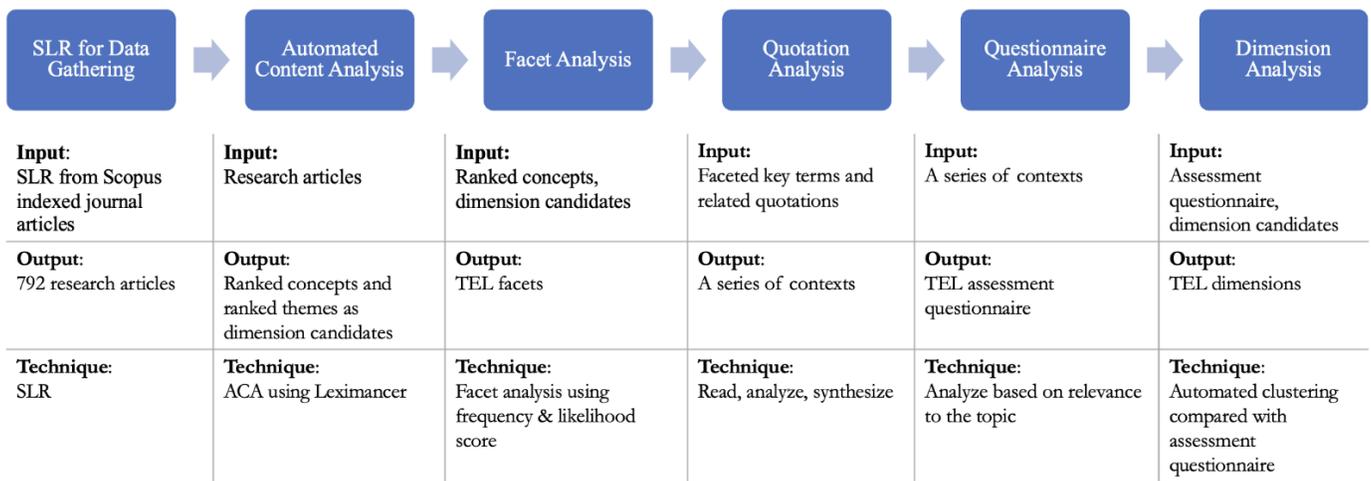


Fig. 2. Methodology with Detailed Input, Output and Techniques used.

B. ACA with Leximancer

Leximancer assists the ACA process by importing data, generating concept seeds, generating a thesaurus, and visualizing the results. The process begins with importing previously selected articles. Then, the tool parses the terms from all articles, weighting them based on both the frequency and likelihood of occurrence. This weighting score becomes a basis for sorting terms, eliminating low-scoring terms, and assigning high-scoring terms as resulting concepts. The data learning process is continuous, starting with selecting a concept seed from the list of terms. A *concept* is a highly relevant term that represents the topic in a collection of research articles as data.

The concept seed is the starting point of the definition of such a concept. The process then adds any highly related terms into the concept’s definition. Thus, if there are more relevant terms than the seed, a new concept seed can be generated. This process continues until all terms have been processed. The term with the highest weighting score becomes a concept that represents the various meanings of the terms. The concepts that are highly related to each other will be clustered into a higher level of data representation, called a theme.

Leximancer will then visualize the formation themes and the sequence of concepts based on the count of hits, as shown in Fig. 3. The illustration depicts the resulting concepts (relevant and meaningful terms) and themes (a group of interrelated concepts). The themes are frequently analyzed as dimension candidates in related works, though they can change according to the theme size setting. However, there is also no exact formula for the optimal size. In addition, some of the generated themes are a collection of outliers, namely concepts that are not closely related to the central concept. Such outliers are inappropriate for this study, which looks for concepts and dimensions capable of representing TEL. In the previous work [19], the selection of candidate dimensions was done only by the highest score concepts.

Selection of candidate dimensions in this research adopts Luhn’s theory, using concepts’ hits of occurrence as the basis for analysis. The related work also used this theory [18] for choosing the key terms (called concept in this research). This study chooses the main themes for candidate dimensions by determining an upper and lower cut. The size setting for the theme is determined with prudence at 25%, considering the condition where the bar chart could present the hits of occurring numbers with a clear distinction between the upper and lower cut, as shown in the Analysis Synopsis tab in Fig. 3. This setting represents a condition where several themes located between the upper and lower cut can be assumed to be the most relevant, themes above the upper cut are considered too general, and those below the lower cut are less relevant. The resulting themes cover “Students,” “Used,” “Study,” “Different,” “Educational,” and “Teacher.” These selected themes will be further examined during facet analysis.

The next step is to break down each selected theme into a list of concepts as the output. The Luhn theory [16] is also used in this process. The concepts are examined based on the terms’ occurrence counts and the likelihood of occurrence related to the theme as the main concept. Fig. 4 and Fig. 5 show graph illustrations that present sorted concepts based on the occurrence likelihood and count. The detailed theme in the figures is Students.

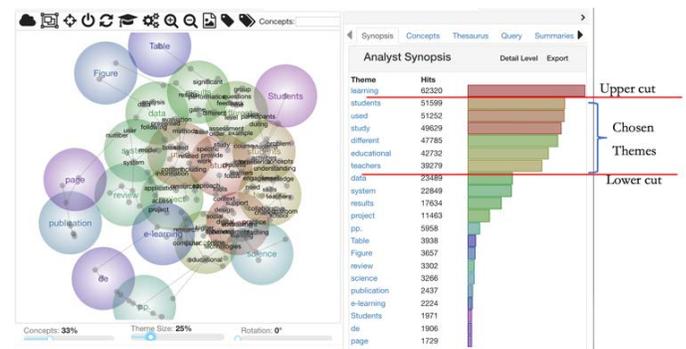


Fig. 3. Resulting Concepts and Themes in Leximancer.

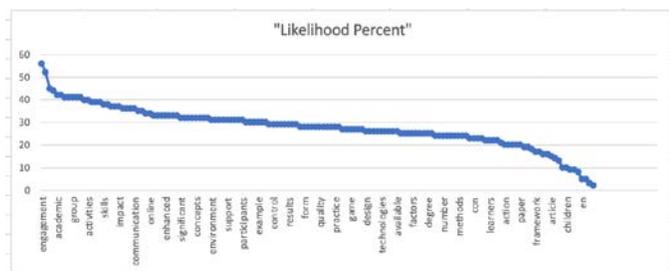


Fig. 4. Graph for Choosing Relevant Concepts in the Students Theme based on Likelihood.

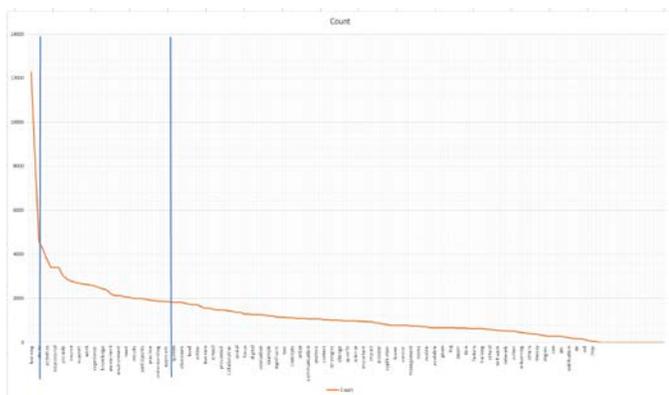


Fig. 5. Graph for Choosing Concepts in the Students Theme based on Count.

The likelihood is the possible percentage of the occurrence between two concepts. In the likelihood graph (Fig. 4), the concepts on the left side lower cut are chosen, which is the condition before the curve is flattened. The count is the frequency of the related concept and the topic of discussion. The chosen concepts have counts located between the upper and lower cut. The lower cut is when the curve starts to flatten, while the upper cut is when the curve starts to decrease, but not too sharply. Next, the resulting list of concepts from Fig. 4 and Fig. 5 are cross-compared to the concepts that occur in both, thus generating a list of concepts for a theme. Each chosen theme is further explored following this approach.

### C. Facet Analysis for Organizing ACA Results

The output of ACA is a collection of the most relevant concepts representing the scope of the TEL discussion. These findings, however, cannot be directly converted into assessment questionnaires. Accordingly, the contribution proposed in this research is how to interpret this list of concepts and translate it into expected results according to the research objective.

The generated concepts serve as the TEL key terms. The next step is to apply facet analysis by examining each concept using the specific criteria. The first criterion, which is listed in the concepts in the topic guide, is one of Leximancer’s outputs, representing a subject index for an extensive document collection. The second is listed concepts that are not topic-specific or highly related to research and writing terminology, such as conjunctions, verbs, or adjectives.

This resulting concept collection is then organized into a logical classification, written as hierarchical structures. It is the performed process to construct TEL facets. Every structure represents a distinct aspect of a story related to the topic under discussion. Several alternative methods exist to develop a facet: drawing from Leximancer’s topic guide as it is (clear description); structuring several concepts into a make-sense facet (need analysis), and digging deeper into sub-concepts to attain the meaning (need deeper analysis). The three alternatives may become a recommendation if weighting is required in the TEL maturity assessment.

The results of the facet analysis for the Students theme can be seen in Fig. 6. The first result, referred to as a “clear description,” is the “problem students” facet, which has a sub-facet covering problems, skills, thinking, and understanding. The “need analysis” is a student facet due to the logically structured various concepts using the meaning of terms. The school facet covers “teachers,” “classroom,” “class,” and “digital” as sub-facets. The last step is incorporating another concept (term) to the chosen concepts to deliver helpful insight, which is part of the “needs deeper analysis.” For example, in the student inquiry facet, the concept of “questions” complements the meaning. This process is repeated for all the themes, resulting in the TEL facets. The next concern is how to obtain useful knowledge from this collection of the TEL facets.

### D. Quotation Analysis to Contextualize TEL Facets

The next step is to investigate the underlying insight from TEL facets by putting them into context. Facet analysis allows us to comprehend the topic of discussion using various aspects. A quotation is used to learn the context of an aspect, which is a phrase or sentence in which the facet occurs. The process searches for quotations that contain the facet, covering a combination of concepts. Table I shows an example of examining the context of a TEL facet by finding several relevant quotations from the Students theme. Each facet may contain more than one aspect and insightful knowledge.

For example, the facet on student problems delivers the context as learning objectives. This context is determined using quotations about learning objectives, student activities in class, and technology to improve learning. The quotation search also employs Leximancer by defining searching query using a combination of concepts with the help of an “AND” operator.

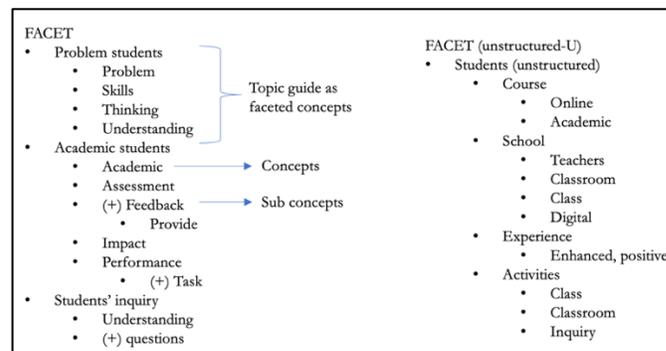


Fig. 6. Graph for Choosing Concepts in the Students Theme based on Frequency.

TABLE I. ILLUSTRATION OF PUTTING CONTEXT INTO TEL FACETS USING QUOTATION ANALYSIS

| No | Facets                                                       | Quotation Examples                                                                                                                                                                                                                                                                                                                                                                                                                      | Context Conclusion                                                             |
|----|--------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| 1  | Problem students (Problems, Skills, Thinking, Understanding) | <ul style="list-style-type: none"> <li>• “Using these scaffolds, and WISE as a meta-context, students constructed and solved problems.”</li> <li>• Help students find or generate their own goals in the problem solving</li> <li>• Students’ learning processes in technology-enhanced ..... and students exercised .... skills and developed understanding .... design thinking, problem-solving, critical thinking, .....</li> </ul> | Problem-solving, exercise skills, thinking, and understanding in class lessons |
| 2  | School (Teachers, Classroom, Class)                          | <ul style="list-style-type: none"> <li>• Every classroom was fitted with an interactive whiteboard; there was a class set of wireless laptops as well as a small ICT suite; and each teacher had their own laptop.</li> <li>• Classroom teachers and students reported new possibilities ... use of digital technologies ... laptops and interactive whiteboards for starting lessons and introducing new tasks.....</li> </ul>         | The use of technology in class                                                 |

This process repeats to analyze each aspect of each chosen theme. The use of concept combination queries, such as using two or three combinations, is also expected. For example: “problem + students” or “student problem” as a compound word, or “problem students + skills,” and so on. This alternative query combination is used to improve the quality of the obtained quotations. After many iterations searching for a facet quotation, the process starts to deliver quotations with the same meaning. The process is reasonable considering that each facet together represents one main topic. As a result, the redundancy elimination process must begin at this point. These collections of concluded contexts became the basis for formulating the TEL questionnaire.

#### E. Interpreting TEL Facets to Formulate Assessment Questionnaires and Dimensions

This advanced step explains how to formulate the TEL assessment questionnaire by understanding the TEL facet’s context, with the illustration shown in Table II. A row corresponds to a facet. Each context is associated with one or more statements. A questionnaire is then designed as a series of statements accompanied by a Likert scale asking how much the respondent agrees with the statement. Each context and the questionnaire item are also scrutinized for its type of respondent target.

This process is repeated until all possible contexts for each candidate dimension have been discussed. Further analysis is performed to compare each question representing each candidate dimension to reduce the possibility of redundant statements. The result is a distributed array of statements for each candidate dimension.

This process repeats until all possible contexts for each theme have been analyzed. Further analysis is performed by comparing each theme’s questionnaire statements to reduce the possibility of redundant statements. As a result, we can finally obtain the complete questionnaire formulation. The resulting statements are naturally clustered into groups based on the dimensions candidates generated in the previous step (subsection B). However, this study conducts more clustering processes to determine additional representable groups with aligned and explicit ideas, called dimensions. These findings are consistent with the grouping of questionnaire statements that have minimized redundancy.

## IV. RESULT AND DISCUSSION

The discussion includes insight about assessment weighting recommendations from the quotation analysis, constructed

questionnaire examples, and the dimensions covering the questionnaire.

#### A. Quotation for Context

In line with the facet determination process, three criteria related to the difficulty of the analysis process also determine the quotation analysis. The process begins by identifying quotations that clearly describe the context, those that require simple analysis, and quotations requiring further analysis. The ease of performing the analysis is related to the context’s relevance to TEL.

The relevance is high if the meaning is clear and explicit, and vice versa. Establishing the quotation may also become a basis for defining the TEL assessment weighting. Thus, every facet has different measure in depicting the data story as follows: clear description: High; needs simple analysis: Medium; requires deeper analysis: Low.

#### B. TEL Maturity Assessment Questionnaires

Table II shows an example of the statement formulation for the Students dimension, with the sample questionnaire for each dimension as delivered in the Appendix. There are several concerning issues, however, in constructing the questionnaire statements. The first is translating the English context into Indonesian Bahasa statements. This study was conducted in Indonesia; therefore, the prospective respondents are Indonesians who are not English natives.

The second issue is a requirement to identify respondents’ demography. Therefore, preliminary questions are needed to help understand the characteristics of the respondents, that is, respondents’ assessment questionnaires. The questions are not formulated from the previously described process but are customized to the needs of the research. As a result, the customized preliminary questions reduce potential confusion when applied to pilot respondents. The information is then used to improve the questionnaire before it is widely distributed.

Concerns are also evident regarding what role is appropriate for a respondent for a specific question. Based on the analysis, it was discovered that the students could not answer all of the questions. As a result, the respondents’ selections became broader; namely, students, teachers, both students and teachers, and institutions. The expanded scope raises the question of how to process the data and regulate the proportion of each respondent’s role in the TEL maturity assessment.

TABLE II. ILLUSTRATION OF HOW TO DEVELOP A TEL MATURITY ASSESSMENT QUESTIONNAIRE

| No | Contexts                                                                                                                                                                                                                                                                   | Target Respondent  | Questionnaire Statements                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1  | Problem-solving, exercise skills, thinking, and understanding in-class lessons                                                                                                                                                                                             | Students, teachers | Existing technology can assist students in comprehending the solutions to problem or questions encountered during a learning activity.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
| 2  | <ul style="list-style-type: none"> <li>Technology to improve academic achievement</li> <li>Technology helps student assessments</li> <li>Technology enables more detailed feedback</li> <li>Technology has both good and adverse effects on student performance</li> </ul> | Students, teachers | <ul style="list-style-type: none"> <li>[student/teacher] Technology can help students improve their academic performance (improvement of grades).</li> <li>[teacher] Existing technology can aid in the process of evaluating student learning outcomes.</li> <li>[student/teacher] Technology has been used to reach all students for them to receive feedback (comments, improvements) on the outcomes of their work on their assignments.</li> <li>[student/teacher] Technology positively impacts student learning outcomes (e.g., quiz/exam scores).</li> <li>[student/teacher] Technology harms student learning outcomes (interference) (example: due to cellphone addiction for less-productive matters)</li> </ul> |

Another concerning issue arises when establishing the context of the collection of quotations. It may be possible that researcher subjectivity is present. However, this study includes efforts to reduce researcher bias and subjectivity while increasing the trustworthiness of the research by conducting quantitative data processing using ACA before the context analysis. This subjectivity may be assumed as part of the researcher’s reasoning when analyzing and interpreting the data.

C. TEL Maturity Assessment Dimensions

The following seven dimensions were determined based on the initial dimension candidates, context adjustment, and regrouping the generated questions with an aligned idea. Each dimension represents influencing factors and being influenced by technological usage. Table III describes the dimensions and the corresponding assessment questionnaire statements. In previous work [19], the result only consists of three general dimensions: technology advancement, improved learning design, student’s achievement.

TABLE III. TEL MATURITY DIMENSIONS

| Dimensions                      | Number of Questions | Descriptions                                                                                   |
|---------------------------------|---------------------|------------------------------------------------------------------------------------------------|
| Student Performance             | 9                   | Relates to improving student performance                                                       |
| Teachers & Teaching             | 9                   | Related to improving the ability of teachers and improving teaching                            |
| Learning Process                | 8                   | Related to improving learning                                                                  |
| Accessible & Applied Technology | 9                   | Related to access to various uses of technology                                                |
| Contents                        | 9                   | Related to content and learning resources                                                      |
| Strategy & Regulation           | 9                   | Related to the formulation of learning strategies using technology and conformity to the rules |
| Technology Governance           | 9                   | Regarding technology, governance to support learning                                           |

The number of questionnaire statements for each dimension is determined by general statistical provisions stating that the number of related questions should not be excessively different. The method assumes that each dimension is equally important in achieving a level of TEL maturity. The following

activity tests the questionnaire on pilot respondents. Thus, the weight of each dimension influencing TEL maturity can be recalculated in future research. Moreover, the formulation of these dimensions can be helpful in the construction of a TEL working framework. Recommendations for improving the TEL maturity assessment results will also be made for each dimension.

V. CONCLUSION AND LIMITATIONS

This study investigates how to use SLR in conjunction with ACA to create TEL maturity assessment instruments constituting questionnaires and dimensions. This research attempts to combine automatization in qualitative data analysis using ACA and qualitative data interpretation using facet analysis. The methodology include data gathering using SLR, data processing using ACA, organizing resulting concepts using facet analysis, searching quotations matching the meaning of TEL facets, inferring contexts from TEL facets, and determining dimensions of TEL maturity.

This research is part of a more extensive study to determine the TEL maturity assessment instrument, which attempts to take a novel approach. The approach used in this study can become a recommendation for how SLR can help novice researchers formulate assessment questionnaires through a literature review and discover how qualitative analysis can be initiated with a quantitative approach to reduce bias and subjectivity of the researchers.

In this study, 64 questionnaire statements were assembled and categorized based on the respondents’ target, either the institution, the student, the teacher, or both. This questionnaire statement set is also grouped based on the TEL maturity constructs, which are referred to a dimension. The seven dimensions are students’ performance, learning process, applied technology, contents, accessibility, teachers and teaching, and strategy and regulation.

This study has some limitations, including a critical point in implementing certain stages. The first obstacle is when the context is reduced to questions in different languages (from English to Indonesian) so that pilot respondents are still required to accept the predetermined questionnaire formulation. The second is the point of view used in developing the questionnaire: a higher education institution in this study. As a result, the findings may not apply to other

educational levels. The third issue is that the focus on which instrument was built in the use of technology primarily from the learner's perspective to improve learning, so the pedagogical aspect is not discussed in depth. The fourth issue is how to justify the influencing portion of each type of respondent on the measurement results.

Future research is required to test the questionnaire on pilot respondents to see if they can understand, conform with the assessment's objective, and examine the correlation between questionnaire statements and dimensions as a mutually exclusive entity. These findings would then be used to develop a maturity appraisal and scoring method, including the weighting to establish the TEL maturity working framework. Furthermore, the questionnaire's development results are used as a reference for developing TEL improvement roadmap.

#### ACKNOWLEDGMENT

This work is funded by Universitas Indonesia through Pendampingan Publikasi Internasional Q2 (PPI Q2), Grant No. NKB-550/UN2.RST/HKP.05.00/2021.

#### REFERENCES

- [1] M. Alrasheedi, "A Maturity Model for Mobile Learning," London, Ontario, Canada, 2015.
- [2] S. Aleem, L. F. Capretz, and F. Ahmed, "A Digital Game Maturity Model (DGMM)," *Entertainment Computing*, vol. 17, pp. 55–73, Nov. 2016, doi: 10.1016/j.entcom.2016.08.004.
- [3] S. Aleem, "A Digital Game Maturity Model," 2016.
- [4] D. Proenca and J. Borbinha, "Enterprise Architecture: A Maturity Model Based on TOGAF ADM," Jul. 2017, pp. 257–266. doi: 10.1109/CBI.2017.38.
- [5] A. Bierema et al., "Quantifying cognitive bias in educational researchers," *International Journal of Research & Method in Education*, vol. 44, no. 4, pp. 395–413, Aug. 2021, doi: 10.1080/1743727X.2020.1804541.
- [6] B. Hyndman and S. Pill, "What's in a concept? A Leximancer text mining analysis of physical literacy across the international literature," *European Physical Education Review*, vol. 24, no. 3, pp. 292–313, Aug. 2018, doi: 10.1177/1356336X17690312.
- [7] X. Lin, H. Zhang, H. Wu, and D. Cui, "Mapping the knowledge development and frontier areas of public risk governance research," *International Journal of Disaster Risk Reduction*, vol. 43, p. 101365, Feb. 2020, doi: 10.1016/j.ijdrr.2019.101365.
- [8] B. Hjørland, "Facet Analysis: The Logical Approach To Knowledge Organization," *Information Processing & Management*, vol. 49, no. 2, pp. 545–557, Mar. 2013, doi: 10.1016/j.ipm.2012.10.001.
- [9] A. Kirkwood and L. Price, "Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? A critical literature review," <https://doi.org/10.1080/17439884.2013.770404>, vol. 39, no. 1, pp. 6–36, 2014, doi: 10.1080/17439884.2013.770404.
- [10] P. Nicoll, S. MacRury, H. C. van Woerden, and K. Smyth, "Evaluation of Technology-Enhanced Learning Programs for Health Care Professionals: Systematic Review," *Journal of Medical Internet Research*, vol. 20, no. 4, p. e131, Apr. 2018, doi: 10.2196/jmir.9085.
- [11] R. Caralli, M. Knight, and A. Montgomery, "Maturity Models 101: A Primer for Applying Maturity Models to Smart Grid Security, Resilience, and Interoperability.," Fort Belvoir, VA, Nov. 2012. doi: 10.21236/ADA610461.
- [12] A. Rahmah, H. B. Santoso, and Z. A. Hasibuan, "E-Learning Process Maturity Level: A Conceptual Framework," in *Journal of Physics: Conference Series*, 2017, vol. 978, p. 12028. doi: 10.1088/1742-6596/978/1/012028.
- [13] B. Kitchenham et al., "Systematic Literature Reviews in Software Engineering - A Tertiary Study," *Information and Software Technology*, vol. 52, no. 8, pp. 792–805, Aug. 2010, doi: 10.1016/j.infsof.2010.03.006.
- [14] B. J. Biroscak, J. E. Scott, J. H. Lindenberger, and C. A. Bryant, "Leximancer Software as a Research Tool for Social Marketers: Application to a Content Analysis," <https://doi.org/10.1177/1524500417700826>, vol. 23, no. 3, pp. 223–231, Apr. 2017. doi: 10.1177/1524500417700826.
- [15] V. Wilk, H. Cripps, A. Capatina, A. Micu, and A. E. Micu, "The state of #digitalentrepreneurship: a big data Leximancer analysis of social media activity," *International Entrepreneurship and Management Journal*, vol. 17, no. 4, pp. 1899–1916, Dec. 2021, doi: 10.1007/S11365-020-00729-Z.
- [16] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM Journal of Research and Development*, vol. 2, no. 2, pp. 159–165, Apr. 1958, doi: 10.1147/rd.22.0159.
- [17] M. Dehghani, H. Azaronyad, J. Kamps, D. Hiemstra, and M. Marx, "Luhn Revisited: Significant Words Language Models," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management - CIKM '16*, 2016, pp. 1301–1310. doi: 10.1145/2983323.2983814.
- [18] A. Rahmah, H. B. Santoso, and Z. A. Hasibuan, "Characteristics analysis for technology enhanced learning maturity: A qualitative approach — Universitas Indonesia," Presented on 27th International Conference on Computers in Education 2019. [http://ilt.nutn.edu.tw/icce2019/04\\_Proceedings.html](http://ilt.nutn.edu.tw/icce2019/04_Proceedings.html) (accessed Dec. 28, 2021).
- [19] A. Rahmah, H. B. Santoso, and Z. A. Hasibuan, "Conceptualizing Technology-Enhanced Learning Constructs: A Journey of Seeking Knowledge using Literature-Based Discovery: In Press," *Computing Conference*, London, 2020.
- [20] A. C. Ferreira, B. C. M. dos S. Maculan, M. M. L. Naves, A. C. Ferreira, B. C. M. dos S. Maculan, and M. M. L. Naves, "Ranganathan and the Faceted Classification Theory," *Transinformação*, vol. 29, no. 3, pp. 279–295, Dec. 2017, doi: 10.1590/2318-08892017000300006.
- [21] M. Usman, R. Britto, J. Börstler, and E. Mendes, "Taxonomies in software engineering: A Systematic mapping study and a revised taxonomy development method," *Information and Software Technology*, vol. 85, pp. 43–59, May 2017, doi: 10.1016/J.INFSOF.2017.01.006.

APPENDIX

This section contains detailed questionnaire statements for each dimension.

APPENDIX. I. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: STUDENT PERFORMANCE

| No | Questionnaire Statement                                                                                                                                         | Respondent      |
|----|-----------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 1  | The existing technology can help students understand problem solutions/answer the questions in a learning activity.                                             | Student/Teacher |
| 2  | The existing technology can help students practice skills and understanding in a learning activity                                                              | Student/Teacher |
| 3  | Technology can help students improve academic achievement (improving grades)                                                                                    | Student/Teacher |
| 4  | Technology does not have a bad influence (distraction) on student learning outcomes (example: addiction due to social media, gaming, or less productive things) | Student/Teacher |
| 5  | The existing technology makes it easier for students to get answers to questions/curiosity/explore things related to learning topics                            | Student/Teacher |
| 6  | The existing technology can help all students, both those who have good and bad grades                                                                          | Student/Teacher |
| 7  | Students have been able to use technology for learning activities.                                                                                              | Student         |
| 8  | The use of technology for learning improves students' digital literacy skills                                                                                   | Student         |
| 9  | Spending more time using technology for learning improves student learning outcomes                                                                             | Student         |

APPENDIX. II. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: TEACHERS AND TEACHING

| No | Questionnaire Statement                                                                                          | Respondent      |
|----|------------------------------------------------------------------------------------------------------------------|-----------------|
| 1  | Existing technology can help facilitate the process of assessing student learning outcomes                       | Teacher         |
| 2  | The teaching process has used technology to support collaboration in group learning                              | Student/Teacher |
| 3  | The learning process has used technology as a medium for assessment, evaluation, and feedback.                   | Student/Teacher |
| 4  | There are adjustments to instructional design that allow for improved learning using technology                  | Teacher         |
| 5  | The successful use of technology requires digital literacy skills in using various applications                  | Student/Teacher |
| 6  | The successful use of technology requires digital literacy skills in utilizing technology to support teaching    | Teacher         |
| 7  | Development of an online learning environment requires knowledge of technology and teaching and learning content | Teacher         |
| 8  | The teacher's role is to facilitate independent learning, not the giver of knowledge                             | Student/Teacher |
| 9  | Teachers have confidence in using technology for teaching                                                        | Teacher         |

APPENDIX. III. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: LEARNING PERFORMANCE

| No | Questionnaire Statement                                                                                                                | Respondent      |
|----|----------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 1  | Technology has been used so that all students can get feedback (comments, improvements) on the results of working on their assignments | Student/Teacher |
| 2  | Existing technology allows students to explore things related to learning topics in several ways                                       | Student/Teacher |
| 3  | Existing technology allows student learning to be carried out in a blended or full online manner.                                      | Student/Teacher |
| 4  | Students enjoy using technology to help with learning activities.                                                                      | Student/Teacher |
| 5  | The learning process has used technology to support the exploration process (inquiry).                                                 | Student/Teacher |
| 6  | Learning methods and technology allow students to learn independently                                                                  | Student/Teacher |
| 7  | All learning activities have used digital technology                                                                                   | Student/Teacher |
| 8  | The use of technology has supported collaboration and collaboration of students in learning and gaining understanding                  | Student/Teacher |

APPENDIX. IV. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: ACCESSIBLE AND APPLIED TECHNOLOGY

| No | Questionnaire Statement                                                                                                                                     | Respondent      |
|----|-------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 1  | Technologies such as laptops, tablets, smartphones, or computers are available to access the online classroom.                                              | Student/Teacher |
| 2  | Technologies such as learning management systems (e.g., Moodle) and MOOCs are available for learning activities.                                            | Student/Teacher |
| 3  | The learning process has used various digital technologies and applications that can help students learn, as well as improve their understanding and skills | Student/Teacher |
| 4  | The learning application used can be accessed via a computer/laptop or cellphone (mobile devices).                                                          | Student/Teacher |

|   |                                                                                                                                            |                 |
|---|--------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 5 | Learning has used a variety of the latest technologies as learning media (such as games, mobile applications, virtual classes, and others) | Student/Teacher |
| 6 | Learners can access a variety of learning technologies that suit their needs                                                               | Student         |
| 7 | The learning process has used technology such as social media as a means of interaction, communication, collaboration in a virtual space   | Student/Teacher |
| 8 | The success of applying technology to improve learning is influenced by an understanding of the technology                                 | Student/Teacher |
| 9 | Good understanding of the use of technology to support learning                                                                            | Student/Teacher |

APPENDIX. V. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: CONTENTS

| No | Questionnaire Statement                                                                                                        | Respondent      |
|----|--------------------------------------------------------------------------------------------------------------------------------|-----------------|
| 1  | Existing technology allows all learning content to be accessed online and openly                                               | Student/Teacher |
| 2  | Technology can help teachers, educational institutions, and students create learning content that can be used for all students | Student/Teacher |
| 3  | Learning content continues to be improved so that students continue to learn with the help of technology                       | Student/Teacher |
| 4  | Learning content (materials, assignments) can be accessed anytime (open, available) by learners                                | Student/Teacher |
| 5  | Learning content and media can be accessed online, anywhere, anytime                                                           | Student/Teacher |
| 6  | Learning content can be accessed anytime via mobile devices                                                                    | Student/Teacher |
| 7  | Easier access to learning resources improves the continuity of the learning process                                            | Student/Teacher |
| 8  | The class has used various learning materials (not only one) to give to students                                               | Student/Teacher |
| 9  | Availability access to online learning content and resources                                                                   | Student/Teacher |

APPENDIX. VI. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: STRATEGY AND REGULATION

| No | Questionnaire Statement                                                                                          | Respondent                   |
|----|------------------------------------------------------------------------------------------------------------------|------------------------------|
| 1  | Existing technology can help students learn and explore both inside and outside the classroom.                   | Student/Teacher              |
| 2  | The learning process uses e-learning as a tool for learning management that can be accessed anywhere             | Student/Teacher/ institution |
| 3  | Utilization of technology supports the learning process following applicable laws and regulations                | Teacher/institution          |
| 4  | Learning and teaching has been student-centered (focused on students)                                            | Student/Teacher//institution |
| 5  | Existing learning has implemented personalized learning for students according to their needs                    | Teacher/institution          |
| 6  | Curriculum and teaching strategies have included technology usage as one of the considerations                   | Teacher/institution          |
| 7  | The rules and procedures in teaching activities have supported the use of technology and its development         | Teacher/institution          |
| 8  | All supporting administrative for learning activities have used digital technology                               | Teacher/institution          |
| 9  | Classes have been able to facilitate students with diverse backgrounds and abilities with the help of technology | Teacher/institution          |

APPENDIX. VII. QUESTIONNAIRE: STATEMENTS FOR DIMENSION: TECHNOLOGY GOVERNANCE

| No | Questionnaire Statement                                                                                                      | Respondent           |
|----|------------------------------------------------------------------------------------------------------------------------------|----------------------|
| 1  | There is teacher participation in determining the development and use of technology for learning                             | Teacher/ Institution |
| 2  | There is a continuous development of technology for learning                                                                 | Teacher/ Institution |
| 3  | The success of applying technology to improve learning is influenced by the duration of technology use                       | Teacher/ Institution |
| 4  | Institutional management supports the development of pedagogy for the improvement of learning using technology               | Teacher/ Institution |
| 5  | Utilization of technology has been aligned with instructional design and teaching strategies                                 | Teacher/ Institution |
| 6  | Teachers have sufficient time to enhance learning with existing technology                                                   | Teacher/ Institution |
| 7  | There has been an evaluation of the learning environment using technology based on learning methods, access, and ease of use | Teacher/ Institution |
| 8  | Technology usage is already based on the need to improve teaching                                                            | Teacher/ Institution |
| 9  | Technology usage promotes cost-effective learning                                                                            | Teacher/ Institution |

# A Knowledge-based Expert System for Supporting Security in Software Engineering Projects

Ahmad Azzazi<sup>1</sup>

Dept. of Software Engineering  
Applied Science Private University, Amman, Jordan

Mohammad Shkoukani<sup>2</sup>

Dept. of Computer Science  
Applied Science Private University, Amman, Jordan

**Abstract**—Building secure software systems requires the intersection between two engineering disciplines, software engineering and security engineering. There is a lack of a defined security mechanism for each of the software development phases, which affects the quality of the software system intensively. In this paper, the authors are proposing a framework to consider the security aspects in all the phases of the software development process from the requirements until the deployment of the software product, with three additional phases that are important to automatically produce a secure system. The framework is developed after analyzing the existing models for secure system development. The key elements of the framework are the addition of the phases like physical, training, and auditing, where they improve the level of security in software engineering projects. The authors found so a solution for the replacement of the knowledge of the security engineer through the construction of an intelligent knowledge-based system, which provides the software developer with the security rules needed in each phase of the software development lifecycle and it improves the awareness of the software developer about the security-related issues in each phase of the software development lifecycle. The framework and the expert system are tested on a variety of software projects, where a significant improvement of security in each phase of the software development process is achieved.

**Keywords**—Knowledge-based systems; security engineering; software development process; expert systems

## I. INTRODUCTION

Software-intensive systems are a major factor in many business areas. There is an increasing need for such software systems that could help us in the daily life. These software systems must fulfil certain requirements. The usage of the internet as a platform for electronic commerce and online banking pushes the need for securely reliable software systems. The software systems must be secured against potential threats and attacks. Attackers are looking for security holes in these software systems to increase their chance of getting sensible information that could be used in an illegal way. As a result of all these facts, the software developers must think about building secure software systems before beginning with the real software systems, and a process of ensuring a secured software system must be followed [1].

There are many disciplines involved in such a process for developing a secure software system, where software engineering is one of the most important disciplines. To develop software in general the authors need a software engineer. The task of such an engineer is to produce quality

software within the given constraints such as time, cost, etc. The software engineer must follow a generalized development process suitable for the desired software [2].

There are many stages in a software development process, the requirement definition, system design, systems implementation, and other stages. In all these different important stages security must be considered and developed in each stage and for each stage. Therefore, there is a need to integrate the security related tasks into each stage of the software development process. The role of a security engineer appears as a part of the software development team. The need for a security engineer or software development enforces the whole development of a software system to interact with different influencers on the final software product [2, 17].

One of the most important dangerous practices during software development is the lack of detailed security requirements. The system requirements must include those ones related to security requirements. The security requirements must be specified and integrated into the whole software development.

The security of existing software majority was built as ad hoc solution that means after the development of the software, security was added to the system, even for security critical systems. The Development of secured software gives us the interaction between two disciplines, software engineering and security engineering. Building secure software is a process in which software security is considered in all phases of a software engineering life cycle [3, 16].

The focus will be on the security activities at each phase of software development, and the authors will present a Knowledge-Based Expert System for Supporting Security in Software Engineering Projects.

The paper is organized as follows: section one is the introduction, section two is the background section which provides a description about software, software engineering, software process, software engineering for Security, security engineering and knowledge-based system for secure software development. Section three is the security software engineering framework using a knowledge-based system provides the proposed framework solution for the research problem stated in this paper. The fourth section includes the Web-Based Security Expert System for Software Engineering Projects. Section five discusses the contribution. Finally, section six is the conclusion.

## II. BACKGROUND

In this section, the authors are giving an overview of the most important terms used in this research. First, the authors are introducing the term software, software engineering essentials, and software engineering for security-critical systems. Secondly, the authors are describing the information security essentials, the common protection mechanisms, and security engineering. Finally, a brief description of the knowledge-based system to support security in the software process is introduced.

### A. Software

Software could be defined as the computer programs, data and documentation which support processes to do an automatic problem-solving task [2].

Due to “pressman,” the software is a vehicle for delivering a product, which supports or directly ides system functionality, controls other programs like an operating system, effects communications like, networking software or helps building other software like, software tools [4].

To deliver software as a product, many properties should be considered, among the most important characteristics is the security of the software as a software system, which could be defined as the system attribute that reflects the ability of the system to protect itself from external attacks that may be accidental or deliberate the resources, the prevention of unauthorized disclosure of information, the extent that the software itself must be hidden or obscured, the trustworthiness of data or resources [3].

The main principal security issues are availability, confidentiality, and integrity. There are other extended properties like non-reputability, accountability, and authenticity.

### B. Software Engineering

Software engineering deals with a detailed production of quality software. There must be a way of organizing the various stages of software development, a process. There is a need for different tools, technologies, and techniques due to the given software problem with the consideration of resources and constraints to be applied. It is the targeted provision and systematic use of principles, methods; process models, concepts, and tools for the development of software systems with a quality focus [4].

### C. Software Engineering for Security

Secure Software development usually requires the engagement and usage of a defined software process which includes the intensive usage of tools, methods, techniques, and technologies. Security errors appear from the lack of a detailed definition of security in the requirements stage, design stage, or coding stage. Therefore, an urgent need for a security engineer appears before going into a stage of the software development process. One must carefully consider the security aspects of the software product from the beginning with requirements and moving on through later lifecycle activities, ending with the deployment and the administration of the software product [5, 15].

### D. Security Engineering

The duties of a security engineer are to ensure the security of software systems during and after the software development process. Security engineering requires cross-disciplinary expertise, ranging from cryptography and computer security through hardware tamper-resistance and formal methods to knowledge of applied psychology, organizational and audit methods [3, 18].

### E. Security in the Software Engineering Life Cycle

Through all the phases of the software engineering phases, a deep consideration of the security aspects must be done for each phase. Beginning with the requirement phase, one should describe the security requirements of the software. In the analysis, a deep analysis of the security requirements should be made. Also, in the design phase, the security design consideration of each designed software component should be considered. In the coding phase, an immense value is given to secure coding and with the deployment phase, all aspects of the security issues must be considered [6].

## III. PROPOSED FRAMEWORK

Knowledge is a justified true belief. Knowledge is a higher level than data or information in a way that it is higher than both, the information is higher than data in its level of abstraction. It is the richest, deepest, and most valuable of the three [7].

There are two kinds of knowledge. The First one is the explicit knowledge, which can be expressed in words and numbers and shared in the form of data, scientific formulae, product specifications, manuals, universal principles, and so forth. This type of knowledge is transmitted in a formal and systematic way among all individuals. Knowledge is another type of knowledge; it is a personal level of knowledge, with difficulties in formulation of it, sharing and communicating it with others. The term knowledge-based stands for the internal structure to process symbols, which represents the information of the real world to achieve intelligent behavior.

The main components of a knowledge-based system are the knowledge base component, the inference component, the user interface, the knowledge acquisition component, and the explanation component [7]. There are diverse types of known knowledge-based systems; the most important types of them like the rule-based systems and expert system, which is a class of software systems, which serves based on expert knowledge to the solution or evaluation of certain problem definitions. Knowledge is represented in diverse ways in expert systems like, the production rules, the semantic networks, and the logic statements representation [7, 14].

Secure Software Systems are associated with a solid software process implementation. Therefore, one must have a security engineer in the software team, who will then guide the software team with security related knowledge through all the phases of the software development life cycle. The lack of security engineers could represent an obstacle in the software development process [8, 13].

Security Engineers could not be available for each software engineering project. In some cases, security engineers are too rare to be found for all the needs of security critical software projects. In other cases, it is too expensive to get extra security engineers into the development team. Also, in some geographical areas it is difficult to find these security engineers. If the security experience of the software development team is not sufficient, this leads to no secure software and the failure of other software. At the end, the software product may completely fail due to the fact of the lack of the security engineers [9, 12].

Therefore, the authors suggest in this work the use of a knowledge-based system to assist the software engineers with the needed security engineering activities. This knowledge-based system is then integrated into the software development framework.

The researchers are proposing a new software engineering framework using some of the knowledge gathered.

Some reasons to choose such a framework are that:

- It covers all aspects of software development life cycle.
- It includes a knowledge base for each phase with the appropriate security related knowledge for each phase.
- It checks the security related activities in each phase.
- The knowledge is adaptive and increased with time.
- Additional phases are easily added to include the security activities in additional phases to ensure more security.
- It is easily implemented and accessed.

The new framework should consider the security aspects in all the phases of the software development process from the specification of the requirements until the construction and after that the deployment of the software product to the security training of the end users.

The framework focuses on the security activities on each phase of the software development process using the general process activities of the software engineering process with knowledge-based system, which helps the developer to get the expert knowledge of a security expert in each phase of the development lifecycle [10, 11].

There are many advantages when using a knowledge-based system (expert system) in the proposed framework like:

- A security knowledge-based system is an intelligent information system, in which security knowledge with methods of knowledge representation and knowledge modeling.
- A knowledge-based system is easy to understand.
- It is more easily to update it according to the increasing level of security knowledge.
- The security expertise feature, where a security expert can make expert level decisions about security.

- The symbolic reasoning feature, where the knowledge is represented symbolically and is given back through a reasoning mechanism.
- The deep security knowledge feature, where a security knowledge base for the different software development lifecycle contains complex security knowledge.
- The security self-knowledge feature, where the system can examine their own reasoning and is able to explain why a specific conclusion is reached.
- They have advantages over conventional systems like; they do not require all initial facts, changes in rules are easily implemented, execution may be done by heuristics or logic, and the effective manipulation of large knowledge.
- They have many benefits when they are used like; increased outputs, increased productivity, decreased decision-making time, increased process, and product quality, reduced downtime, capture of scarce expertise and flexibility.

There are some limitations when using expert system like; that the knowledge is not always readily available, it is sometime difficult to extract expertise from humans, where we have different approaches for the knowledge extraction, and one faces lack of end user trust when using the expert system. Therefore, the researchers have divided the knowledge-based system into many knowledge-based systems for each phase of the different phases of the software development lifecycle to make it easier to get new rules for the expert system extracted from the different sentences of the security expertise.

The proposed framework consists of several steps leading to the development of secured software as shown in Fig. 1.

The Framework begins with the requirement phase, which is specially designed for secured systems and in which the Framework collects the security engineering activities needed for obtaining the security requirements of the software products.

After obtaining the security requirements, the system performs a security test of obtained requirements. The analysis phase of the previous phases follows with the conjunction of the security activities designed for this phase with feedback to the requirements phase to make changes on the requirements of any new requirements that arise in the analysis phase. After it the system begins with the design phase for building the secured product, which has its own security engineering activities designed for this phase. The authors have then a special test for the security validation of the design phase. After that the authors begin with the coding phase with the consideration of the security engineering activities for this phase. The next step in the proposed framework is the security testing of the code that the authors have constructed in the previous step with feedback to the coding phase, to repair any security holes in the code of the product. In the next step, the authors have the deployment of the security-critical system with the security engineering activities that must be done for this phase. The last step in the framework is the security test for the software after the deployment; it includes the feedback

to the deployment phase for making any changes regarding the secure deployment of the software product.

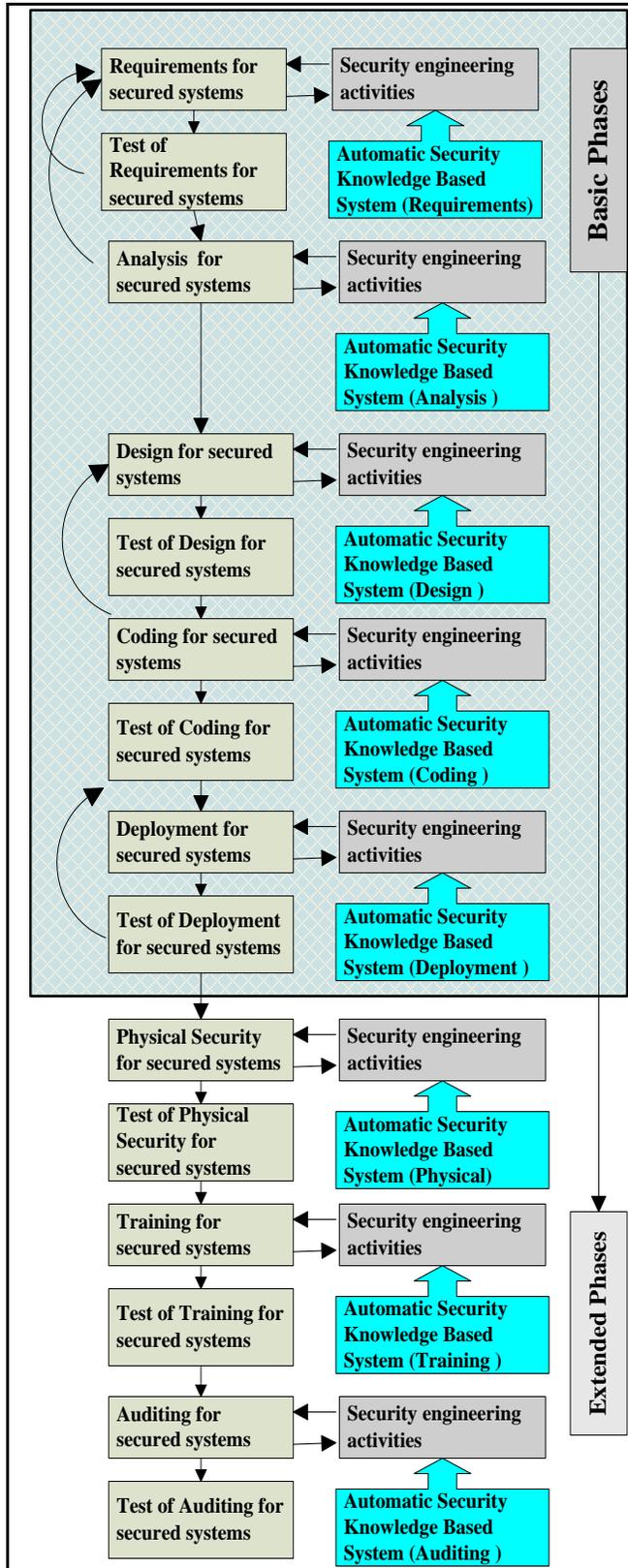


Fig. 1. Security Software Engineering Framework using Knowledge-based System.

The proposed framework is a framework that concentrates on highlighting the security activities to be done in each phase of the general software development lifecycle. It includes an intensive testing of security after each phase. All that leads to the development of a highly secure software product.

The Framework is simple; it could be followed easily from a software engineering perspective. The use of a knowledge-based system assists the software engineers with the needed security engineering activities in different phases of the software development lifecycle.

As shown in Fig. 1 the authors gain the security engineering activities in the requirement phase through an automatic security knowledge-based system especially designed for the requirements phase. For the design phase of the software process lifecycle, the authors have another automatic knowledge-based system for the security activities in this phase.

For the coding phase the authors have also another automatic knowledge-based system as a source for the security engineering activities. For the deployment phase the authors have an automatic knowledge bases system for the security engineering activities as a source. The complicated work the security engineer is done through the usage of different automatic intelligent knowledge-based systems.

Three additional phases are added to the general phases of the software engineering lifecycle, the physical security phase, where it is concerned about the physical security of every essential asset of the whole system, the security training phase, where it is concerned about receiving appropriate information about security training in a software project and the security auditing phase, where it is concerned about involving auditing and monitoring activities of the security requirements of the whole system.

Some limitations of the proposed framework are:

- The availability of knowledge in each phase.
- The changeability of the knowledge over time.
- The amount of knowledge which can increase dramatically.

#### IV. WEB-BASED SECURITY EXPERT SYSTEM FOR SOFTWARE ENGINEERING PROJECTS

In this research, the authors tried to find a suitable solution for security in software projects, where most of the existing software, even software for security-critical systems, has been built and is being built in an ad hoc, unsystematic fashion. In this work, the researchers are representing the interaction between the software engineering discipline and the security engineering discipline. The researchers have made a deep study of the concepts from scientific literature, gathered the knowledge needed to propose a new software engineering framework for security critical systems.

In the proposed framework the researchers are considering the security aspects in all the phases of the software development process from the specification of the requirements until the construction and after that the

deployment of the software product. This consideration of all the security activities is not an ad hoc solution to the software product. In the proposed framework security issues are implemented more efficiently. In the proposed framework the researchers considered all the phases of the general software engineering activities with the three additional for security important phases, the physical phase, the training phase, and the auditing phase. In all these phases, the researchers build up an appropriate knowledge base system (expert system) that should help the developers in applying the needed security engineering activities simple. The separation of the knowledge base system into knowledge base systems for each phase of the framework is made, so that the addition of new rules is specially done for each phase separately, which is simpler to do and requires minimal knowledge of security engineering knowledge.

To prove the need of the proposed framework the researchers conducted a deep survey, which provided us with more information about the security implementation issues in the software project.

This survey helped us in developing rules for each phase of the software development life cycle for the security knowledge-based system. This survey is a sample for descriptive analysis and hypotheses testing, to set up conclusions and recommendations about the usability of the proposed framework.

To prove the proposed framework, the researchers constructed an expert system case tool, the Web-Based Security Expert System for Software Engineering Projects (WB-SES-SEP). The construction of the distinct phases of the overall proposed framework is done in a web environment for many important named reasons. In the (WB-SES-SEP) one could select each phase of the 8 phases of the proposed framework, where one could begin with each phase, or the user could easily add new security rules to the knowledge-based system of each phase.

## V. CONTRIBUTION

The main contribution of this research is the new framework for secure software development using a knowledge-based system, where the modeling of security activity rules on each phase of the software development process using the general process activities of the software engineering process in done. In this framework the researchers added three additional phases, which are essential to get better secured software through the software development lifecycle. These three additional phases are not common in the software development lifecycles. The Framework is simple; it could be followed easily from a software engineering perspective. The use of a knowledge-based system assists the software engineers with the needed security engineering activities in different phases of the software development lifecycle, which solves the problem of the availability of security experts at software engineering projects.

The researchers found that the proposed framework consists of several steps leading to the development of better secured software, through the implementing of a case study,

where the proposed framework is simply used through the case tool, which we built for this work.

Another contribution of the work is the results of the deep survey, which provided us with important information about the security implementation issues in software projects like:

- Most of the projects have security objectives, but the integration of security to the product is on average done.
- Only on average of the projects was the notice taken that security must be manageable.
- Very few of the project members have enough exposure to principles and techniques of secure application development.
- Security is considered and implemented as an ad hoc solution.
- Only very few of the projects have a person responsible for reviewing security.
- The security rules are very weak implemented in all the phases of the general software engineering phases.
- The availability of any information security policy is very weak.
- The physical security of the software project is rarely implemented.
- The security related training of people is not done in a good way.

## VI. CONCLUSION

The researchers proposed a new framework for secure software development using a knowledge-based system, where the modeling of security activity rules on each phase of the software development process using the general process activities of the software engineering process in done. In this framework the researchers added three additional phases, which are essential to get better secured software through the software development lifecycle.

The researchers conducted an analytical survey and the test of the framework with the own built case tool led to the following features of the framework:

- One constructs better secured software when following the proposed framework.
- The security knowledge-based system for each phase of the framework consists of all the needed security rules for each phase.
- The framework is easily understandable and easily used. It is simple implemented and followed.
- Through the analysis the researchers gained very important security rules for each phase of the software development lifecycle and the additional cycles of the framework.
- New rules could be added very simply to the existing rules.

- The inference engine of the knowledge-based system, in which the researchers built, contains very accurate decision tables with a rule importance and a rule implementation level.
- The inference engine of the knowledge base system could decide about any implementation level of the rules in each phase separately, or about all the rules in the phase or about the security implementation level of the phase itself.
- With the case tool, the (WB-SES-SEP), the developer can gain experience about the rules to be implemented in a specific phase and one can repeat the processing of the phase rules until he reaches a satisfied level of security for the phase very easily.
- With the case tool, the (WB-SES-SEP), the developer can get a graphical analysis of the current implementation of all security rules in a specific phase very easily.
- The framework is easy assessing the software developer with the needed security engineering activities through a web-based interface accessible globally and all the time.
- The added phases to general software development life cycle phases are improving the level of security in a software engineering project.
- The case tool, the (WB-SES-SEP), could be used as training tool for software developer in a security critical software project so that, they could identify certain security related problems that face the project members during each phase of the software development lifecycle.
- The case tool, the (WB-SES-SEP), could make the project members more concerned about security by displaying the needed security activities needed in each software engineering phase.

#### ACKNOWLEDGMENT

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research.

#### REFERENCES

- [1] M. Mirakhorli, M. Galster & L. Williams, Understanding Software Security from Design to Deployment, ACM SIGSOFT Software Engineering Notes, Volume 45, Issue 2, 2020, pp 25–26.
- [2] Sommerville, I., Software Engineering, 10th Edition, Pearson India, 2018.
- [3] Jens Bürger et al, A framework for semi-automated co-evolution of security knowledge and system models, J. of Systems and Software, Volume 139, May 2018, Pages 142-160.
- [4] Pressman, R.S. Software Engineering: A Practitioner's Perspective, kindle edition, McGraw-Hill, New York, 2019.
- [5] Lada Gonchar & Lada Gonchar, Implementation of Secure Software Development Lifecycle in a Large Software Development Organization, Proceedings of the 21st International Workshop on Computer Science and Information Technologies, 2019.
- [6] R. Matulevičius, Fundamentals of secure system modelling, 1st ed, Springer, 2017.
- [7] Anthony J Rhem, Knowledge Management in Practice, Auerbach Publications; 1st edition, 2016.
- [8] Ross J. Anderson, Security Engineering, second edition, Wiley, 2008.
- [9] A. Johanson and W. Hasselbring, Software Engineering for Computational Science: Past, Present, Future, Computing in Science & Engineering, vol. 20, no. 2, pp. 90-109, 2018.
- [10] Ahmad AlAzzazi, Asim El Sheikh, Security Software Engineering: Do it the right way, Proceeding of the 6th WSEAS International Conference on SIGNAL PROCESSING, ROBOTICS and AUTOMATION, Corfu Island, Greece, 2007.
- [11] Ahmad AlAzzazi, Asim El Sheikh, Security Software Engineering with a Knowledge Based Engineering, WSEAS TRANSACTIONS ON COMPUTER RESEARCH, 2007, pp.276-282.
- [12] Riad, ABM Kamrul, et al. "Plugin-based Tool for Teaching Secure Mobile Application Development." INFORMATION SYSTEMS EDUCATION JOURNAL 19.2 (2021): 2.
- [13] Yurin, Aleksandr Yu, and Nikita O. Dorodnykh. "Personal knowledge base designer: Software for expert systems prototyping." SoftwareX 11 (2020): 100411.
- [14] Wang, Yingxu, and Omar A. Zatarain. "Design and implementation of a knowledge base for machine knowledge learning." 2018 IEEE 17th International Conference on Cognitive Informatics & Cognitive Computing (ICCI\* CC). IEEE, 2018.
- [15] Burnashev, R. A., Ismail Amer, and A. I. Enikeev. "Expert system building tools based on dynamically updated knowledge." Journal of Physics: Conference Series. Vol. 1352. No. 1. IOP Publishing, 2019.
- [16] Burnashev, Rustam A., et al. "Research on the Development of Expert Systems Using Artificial Intelligence." International Conference on Information Systems Architecture and Technology. Springer, Cham, 2019.
- [17] Alguliyev, Rasim, Yadigar Imamverdiyev, and Lyudmila Sukhostat. "Cyber-physical systems and their security issues." Computers in Industry 100 (2018): 212-223.
- [18] Sultan, Sari, Imtiaz Ahmad, and Tassos Dimitriou. "Container security: Issues, challenges, and the road ahead." IEEE Access 7 (2019): 52976-52996.

# Performance Comparison between Lab-VIEW and MATLAB on Feature Matching-based Speech Recognition System

Edita Rosana Widasari<sup>1</sup>, Barlian Henryranu Prasatio<sup>2</sup>, Dian Eka Ratnawati<sup>3</sup>

Dept. of Computer Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia<sup>1,2</sup>

Dept. of Informatics Engineering, Faculty of Computer Science, Brawijaya University, Malang, Indonesia<sup>3</sup>

**Abstract**—Speech recognition systems are widely used for smart applications. The smart application-based speech recognition system has different requirements for processing the human voice. The most common performance in the speech recognition system is essential to observe, since it is necessary to design smart application-based speech recognition systems for people's needs. Moreover, feature matching is the principal part of speech recognition systems since it plays a key role to authenticate, separate one human voice from another, and their different articulation. Therefore, this work proposes a performance comparison of speech recognition systems based on feature matching using Lab-VIEW and MATLAB. The feature extraction involves calculation of Mel Frequency Cepstral Coefficients (MFCC) for each frame. For the matching process, the system was tested 100 times for each five speeches by making changes in articulation with the same vocal cords. This matching process uses DTW (Dynamic Time Warping), and then the testing is based on the most common performance in the speech recognition system's comparison between Lab-VIEW and MATLAB such as accuracy rate, execution time, and CPU usage. Based on experimental results, the average accuracy rate of MATLAB is better than Lab-VIEW. The execution time testing indicates that Lab-VIEW has a shorter execution time than MATLAB. On the other hand, Lab-VIEW and MATLAB have almost the same CPU usage. This result indicates that the performance comparison is able to be used according to the requirements of smart application-based speech recognition systems.

**Keywords**—Speech; articulation; feature matching; Lab-VIEW; MATLAB

## I. INTRODUCTION

In general, the instrument for creating the human voice can be confined into three parts, i.e., lungs, vocal strings, and verbalization [1]. The combination of the vocal strings with verbalization can create an assortment of discourse. The discourse quality, counting unforgiving, tense, breathy, or whispery voice, can be influenced by emotion and temperament [2]. In the last decade, there has been an automated method of identifying words spoken by the human voice and converting them into readable text. This automated method is called speech recognition. Furthermore, the human voice can also be utilized in computer technology by using the speech recognition system.

Speech recognition is utilized to change over talked shape into content to help people needs [3]. The speech recognition

system is widely used for smart applications, e.g., intelligent wheelchair, Google assistant, Alexa, Cortana, Siri, and home assistant [4], [5]. Each smart application-based speech recognition system has different requirements for processing the human voice. Typically, the speech recognition system works through four stages, i.e., speech analysis, feature extraction, modeling, and testing techniques (matching process) [6]. In the speech analysis stage, speech data contains different types of information that appear a human voice identity. The next stage is feature extraction, which takes on features that might be used to match the digital signal of the human voice to a particular pattern. Then, the modeling stage is used to generate speaker models using speaker-specific feature vectors. In the last stage, the speech-recognition system matches a detected word to a known word using testing techniques (matching process). Feature matching is the most important stage of speech recognition since it plays an important role in authenticating and separating one human voice from another and their different articulation. The matching results then identify similarities [7].

In the last decade, some works have observed the performance of speech recognition system separately [8], [9], [10], [11]. However, since the most common performance of speech recognition systems based on feature matching was not observed at all. Thus, it is needed to observe of all the most common performance of speech recognition systems based on feature matching.

To this end, this work proposes a system that can identify a speech using features matching the most common performance of speech recognition systems. The work is aimed to recognize the speech built in two programming languages Lab-VIEW and MATLAB. They were selected since they have been widely used for designing smart application-based speech recognition systems [12], [13]. In Lab-VIEW programming, the structure of the graphical piece of a program chooses the execution stream in which the computer program design interfacing center points by drawing wires [14]. Furthermore, Lab-VIEW programming is able to combine the virtual instrumentation technology and speech recognition system; and also provided password authentication [15]. On the other hand, MATLAB permits framework control, plotting of a work and information and calculation usage. In spite of the fact that numerically nuanced, a tool compartment that employments a typical machine permits get to the computer logarithmic capabilities [16]. Moreover, in

order to address the analysis and testing issue an appropriate software tool is developed using MATLAB that enabled unified framework for tracking the performance of all necessary functions of speech recognition system [17]. These programming language performances were compared when identifying speech to determine the parameters used for processing human voice.

This work focuses on the method of how the system can recognize the speech based on previously stored voice features sequence as the reference signal. This work implemented the voice recognition based on feature matching with changing articulations. Structurally, the speech recognition system requires a dataset used to train the system's sensitivity to speech patterns. In this work, the stored training data is called the dictionary. This dictionary is used as the database in matching the spoken word.

This work organizes the rest of this paper as follows. Section II presents a review of related works. Section III describes the detail process of feature matching. In Section IV express the materials and system design includes the detail description of the data used in this work and the design of proposed system. Section V presents the result and discussion of the performance comparison between Lab-VIEW and MATLAB on feature matching-based speech recognition system. Finally, Section VI is dedicated to conclusion and further works.

## II. RELATED WORK

The speech recognition system is widely used for smart applications. A study [18] estimated that speech recognition needs to achieve close to a 90% accuracy rate for designing smart homes assistants. One of the factors needed in its design is data communication between devices so as to provide security and convenience that meet people's needs [19]. Moreover, the slower or faster time execution of an intelligent wheelchair is according to their current resistance [20]. Afterwards, the voice assistant technologies (such as Google

Assistant, Alexa, Cortana, Siri, etc.) require addressing restrictions like CPU and memory limitations [21]. This is to achieve an efficient on-device streaming speech recognition system. Thus, this work considers that the accuracy rate, time execution, and CPU usage could represent the most common performance in the speech recognition system. Therefore, the consideration of these parameters is important to observe.

According to A. A. A. Zamil, et al. [8], the extricated highlights of the obscure discourse and after that compared them to the put away extricated highlights for each diverse speaker in arrange to distinguish the obscure speaker using a voting mechanism. However, the key process of selecting the extracted features is minimizing the difficulty of speech recognition system computing for matching processes [9]. Therefore, another study has observed the performance of speech recognition system computing [10], [11]. Nevertheless, the most common performance of speech recognition systems based on feature matching was not observed at all. Therefore, the observation of all the most common performance of speech recognition systems based on feature matching (i.e., the accuracy rate, time execution, and CPU usage) is needed.

## III. FEATURE MATCHING

Before the recognition process, pre-emphasis was applied to the voice signal [22]. Pre-emphasis was aimed to suppress high-frequency parts during the production mechanism of the human voice before performing the framing and windowing process. The framing and windowing process was intended to split the speech signal into smaller parts [23] because it needs to be assumed as a stationary signal. Finally, energy detection was performed on each frame to detect the existence of pronunciation in that frame [24].

Structurally, speech recognition consists of four stages: (1) speech analysis, (2) feature extraction, (3) modeling, and (4) testing techniques or matching process. The block diagram of feature matching, which includes all stages of speech recognition, is shown in Fig. 1.

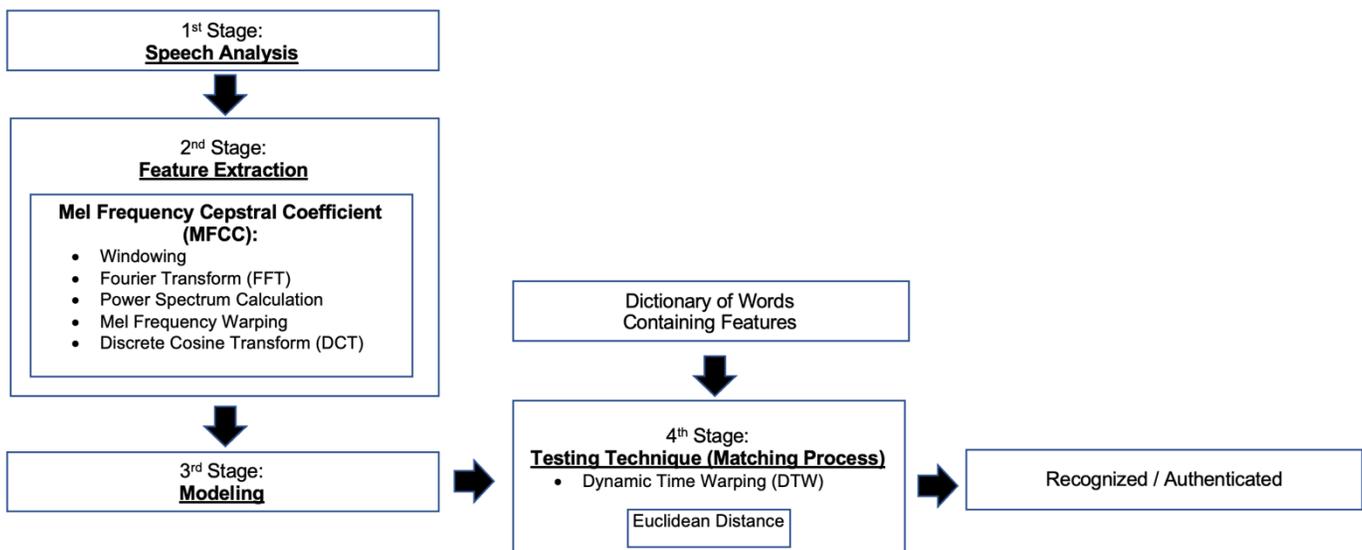


Fig. 1. Block Diagram of Feature Matching.

The first stage is speech analysis, which contains different types of information that shows a human voice identity. After the pronunciation has been detected, this work needs to extract the speech feature on each frame in the second stage. The feature extraction involves calculation of Mel Frequency Cepstral Coefficients (MFCC) for each frame. By using the MFCC, it is able to develop the features from the speech signal which can be used for speech recognition system [25]. The MFCC consist of the following steps: Windowing, Fast Fourier Transform (FFT), power spectrum calculation, Mel Frequency warping, and Discrete Cosine Transform (DCT). The block diagram of the feature extraction process of MFCC is shown in Fig. 2.

Since the effect of frame blocking the speech signal becomes discontinuity, windowing is required in the first process of second stage. Then, FFT is applied to transform speech signal to frequency domain in each frame. Furthermore, the power spectrum for each frame is calculated. However, it is having a lot information which is not needed for feature matching process. Thus, Mel Frequency Warping is used for filtering in the form of a filter bank to determine the size of the power spectrum of a certain frequency band and convert the frequency into mel. Finally, DCT is used for producing a mel spectrum to improve recognition quality [25], [26].

In the third stage, the feature extraction process was done on a set of words and then stored the feature vector sequences [27] as the dictionary. The dictionary function performs feature extraction, which will be stored and used as a reference. The dictionary will be used as a matching reference with the recorded voice speech features. Thus, in the early stages, this work needs to save the voice that will be used as a dictionary dataset. The data were stored in the dictionary and formatted as an array. There were 25 feature vector data sets in the dictionary, which consisted of voice left (5 sets), right (5 sets), up (5 sets), down (5 sets) and stop (5 sets).

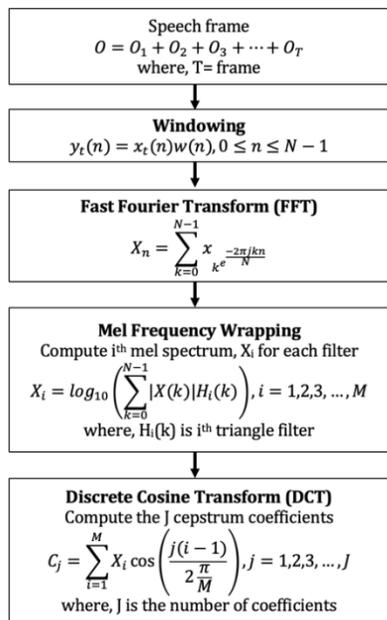


Fig. 2. The Flowchart of Feature Extraction Process of MFCC.

Typically, people have different speaking speeds and characteristics [28]. Dynamic Time Warping (DTW) is used to normalize these differences. Moreover, feature vectors of the voice test sequence were also compared with each word in the dictionary set using the DTW algorithm and the best match in each set was outputted in the final stage. DTW is an algorithm used to measure similarity between two sequences which may vary in time or speed. Thus, it can find the best alignment between two different sequences of signals. Fig. 3 shows the flowchart to implement the DTW algorithm.

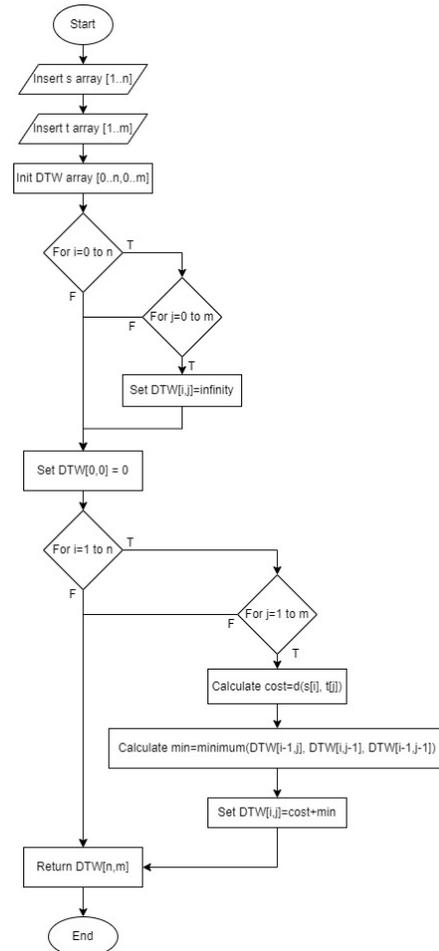


Fig. 3. The Flowchart of DTW Algorithm.

The DTW algorithm works mainly by calculating the Euclidean distance between two points [8], which is the point of the test and reference points for recognizing or authenticating each frame [29]. The threshold was set so that the random noise signal is not generated by the matching speech. Therefore, the DTW algorithm can also be used to find the best matching between voice data test and dictionary data. Then the Euclidean distance function is called when calculating the DTW. After that the matching process between the dictionary voice features with the voice data that has been recorded is performed. Fig. 4 shows the flowchart for calculating Euclidean distance. Finally, this distance will use for comparing voice and reference voice to recognize or authentication process. The detail process of recognize or authentication is described in Section IV below.

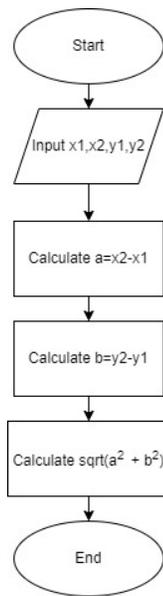


Fig. 4. The Flowchart of Euclidean Distance Calculation.

#### IV. MATERIAL AND SYSTEM DESIGN

This work used the TIMIT Acoustic-Phonetic Continuous Speech Corpus dataset. This is a standard dataset used for evaluation of automatic speech recognition systems. It consists of recordings of 630 speakers of 8 dialects of American English each reading 10 phonetically-rich sentences. It also comes with the word and phone-level transcriptions of the speech. Furthermore, to obtain the work’s objective, the five sets of voice dictionaries (such as left, right, up, down, and stop) is used for matching processes by making changes in articulation.

The system design consists of three parts: Dictionary, DTW, and Euclidean Distance. The Dictionary was used to store five words: left, right, up, down and stop. Each word was repeated 5 times. The voice test sequence will be matched to all words in the first dictionary set and then the word with the least distance will be extracted. Then, the voice test sequence will be matched to all words in the second dictionary set and the matched word will be extracted. The same will be done for the next three sets. Thus, the dictionary will give 5 matched outputs from the dictionary set. The most repeated word in the output will be regarded as the best match.

The DTW can normalize the speed difference of speaking [30]. It can find best-alignment between two different signals. This function will call the Euclidean distance function to calculate the distance between the test signals with reference signals. After that, the results of feature extraction were matched with the dictionary that has been recorded as the reference. This function will do the matching features by calling DTW to find the best alignment between the two sets of sequences.

The voice received by the system compared to reference voice that has been stored in the dictionary. The distance of both speeches was calculated using the Euclidean Distance algorithm. The closest distance to the set reference speech is output for recognizing or authentication process.

#### V. RESULT AND DISCUSSION

In this section, the block diagram and flowchart in the previous section have been implemented in the programming language using Lab-VIEW and MATLAB on a desktop of Intel i7 4-core CPU and 8 GB RAM. Furthermore, this work describes the performance comparison results in the speech recognition system.

##### A. Experimental Result

The testing phase was performed by connecting a microphone to a computer that had installed Lab-VIEW and MATLAB program code. Then, while the speaker speaks a voice, the signals were acquired by the Lab-VIEW for 2 seconds duration at 11025 Hz sampling rate. This work tested the system using 5 different voice speeches consisting of “Left”, “Right”, “Up”, “Down”, and “Stop”. Each of these words was repeated 100 times with different articulations. The detailed information about the amount of data for the matching process can be seen in Table I.

The recognized outputs were observed to calculate the system performance. The first test is the success level of the system in identifying what is spoken. The details of the first testing result of accuracy rate can be seen in Table II.

The average accuracy rate for Lab-VIEW and MATLAB are: 85,6 % and 89,6 %. The accuracy results are obtained by ratio between true positive (TP) and true negative (TN) to TP, TN, false positive (FP), and false-negative (FN) as shown in (1) [31]. TP presents the number of voices that are labeled correctly and TN for the number of voices that are correctly identified as not corresponding to the words spoken. FP indicates the number of voices that are incorrectly labeled. FN denotes the number of voices that are unidentified in words spoken.

TABLE I. DETAILED INFORMATION ABOUT DATA FOR MATCHING PROCESS

| Word         | The number of data |
|--------------|--------------------|
| Left         | 100                |
| Right        | 100                |
| Up           | 100                |
| Down         | 100                |
| Stop         | 100                |
| <b>Total</b> | <b>500</b>         |

TABLE II. THE TESTING RESULT OF ACCURACY RATE IN THE SPEECH RECOGNITION SYSTEM

| Word           | Lab-VIEW Accuracy Rate (%) | MATLAB Accuracy Rate (%) |
|----------------|----------------------------|--------------------------|
| Left           | 81                         | 91                       |
| Right          | 85                         | 85                       |
| Up             | 88                         | 87                       |
| Down           | 85                         | 92                       |
| Stop           | 89                         | 93                       |
| <b>Average</b> | <b>85,6</b>                | <b>89,6</b>              |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

The second test is to test the running time required to execute the program. The execution time is calculated from the voice received by the system until the system finishes identifying the voice. The testing is done 100 times on each word. The second test result of execution time can be seen in Table III, which shows that the average execution time for Lab-VIEW and MATLAB are: 723 ms and 969,4 ms.

TABLE III. THE EXECUTION TIME RESULT OF SPEECH RECOGNITION SYSTEM

| Word           | Lab-VIEW Execution Time (ms) | MATLAB Execution Time (ms) |
|----------------|------------------------------|----------------------------|
| Left           | 752                          | 1044                       |
| Right          | 883                          | 1091                       |
| Up             | 617                          | 953                        |
| Down           | 698                          | 847                        |
| Stop           | 665                          | 912                        |
| <b>Average</b> | <b>723</b>                   | <b>969,4</b>               |

The third test is CPU usage. This test is quite important in building the system so that resources can be used optimally. The testing is done 100 times on each word. The third test result of CPU usage can be seen in Table IV, which shows that the average of CPU usage for Lab-VIEW and MATLAB are: 1,55 % and 1,57 %.

TABLE IV. THE CPU USAGE RESULTS OF SPEECH RECOGNITION SYSTEM

| Word           | Lab-VIEW CPU Usage (%) | MATLAB CPU Usage (%) |
|----------------|------------------------|----------------------|
| Left           | 1,56                   | 1,41                 |
| Right          | 1,84                   | 1,42                 |
| Up             | 1,41                   | 1,74                 |
| Down           | 1,63                   | 1,67                 |
| Stop           | 1,29                   | 1,62                 |
| <b>Average</b> | <b>1,55</b>            | <b>1,57</b>          |

### B. Discussion

Based on experimental results, the average accuracy rate of MATLAB is better than Lab-VIEW since the value of TP and TN in the MATLAB results is higher than Lab-VIEW. On the other hand, Lab-VIEW has a shorter execution time than MATLAB since Lab-VIEW has simple code that is assessed according to the number of elements or lines used. The faster the code can be updated or debugged the better surveyable the program is [32].

In order to verify the performance comparison results, a box-plot method or also called Box-and-Whisker plot method was conducted as presented in Fig. 5. From the figures, this work can find that average accuracy rate and execution time are significantly different between Lab-VIEW and MATLAB using t-test at the level of significance ( $\rho$ ) of 0.05. This result indicates that the average accuracy rate of MATLAB is actually better than Lab-VIEW. Then, Lab-VIEW actually has a shorter execution time than MATLAB since both of them

are significantly different. It is also in accordance with the experimental results.

CPU usage becomes an essential metric to determine how well an application is using the cores. CPU usage refers to a program's usage of processing resources, or the amount of work handled by a CPU. Since the amount and type of managed computing tasks are the same, Lab-VIEW and MATLAB have almost the same CPU usage. Based on the statistical analysis, the CPU usage is not significantly different between Lab-VIEW and MATLAB using t-test at the level of significance ( $\rho$ ) of 0.05. This result indicates that the Lab-VIEW and MATLAB have the same CPU usage since both of them are not significantly different and in accordance with the experimental result.

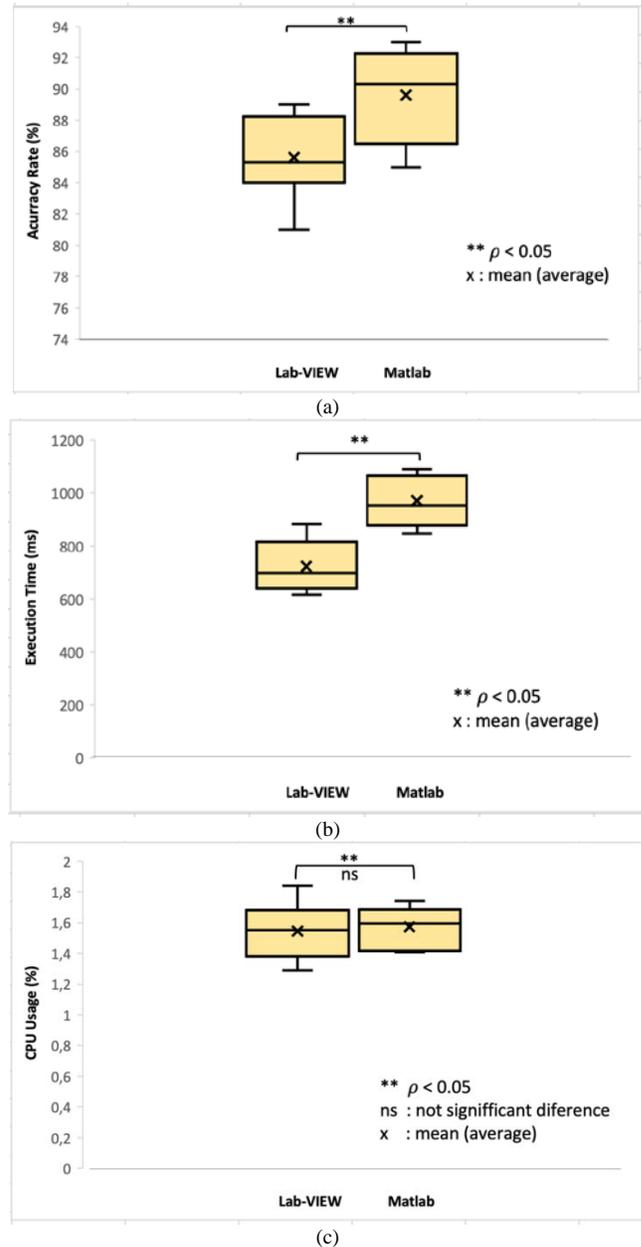


Fig. 5. Box-and-Whisker Plots of Average (a) Accuracy Rate, (b) Execution Time and (c) CPU usage.

## VI. CONCLUSION

This work has designed a speech recognition system using feature matching on Lab-VIEW and MATLAB. The feature was extracted using MFCC and calculating distance using DTW from the speech. Before performing the speech, this work saved a set of the features from speech voice as a training set in a dictionary. The matching process was performed between the feature of the voice and the feature which had been saved in the dictionary. In the testing phase, this work tested five speech words and each word was repeated 100 times. The system experimented using the most common performance in the speech recognition system i.e., accuracy rate, execution time, and CPU usage. The performance comparison results show that the average accuracy rate of Lab-VIEW is 85.6% and MATLAB is 89.6%. The execution time testing of Lab-VIEW is 723 ms and MATLAB is 969.4 ms. While, the Lab-VIEW and MATLAB have almost the same CPU usage which is around 1.5%.

Speech recognition has wide smart applications and includes voice-controlled appliances fully featured speech-to-text software, automation of operator-assisted services and voice recognition aids. This work's result indicates that the performance comparison is able to be used according to the requirements of smart application-based speech recognition systems. Hence, the performance comparison results in improving many speech recognition system applications can be further extended, which can make the process more robust and effective.

## REFERENCES

- [1] Z. Zhang, "Mechanics of human voice production and control," *J. Acoust. Soc. Am.*, vol. 140, no. 4, pp. 2614–2635, 2016. DOI: 10.1121/1.4964509.
- [2] S. Bhatt, A. Jain, and A. Dev, "Acoustic Modeling in Speech Recognition: A Systematic Review," *Int. J. Adv. Comput. Sci. Appl. Sci. Inf. Organ.*, 2020. DOI: 10.14569/IJACSA.2020.0110455.
- [3] S. Yang, "Listener's ratings and acoustic analyses of voice qualities associated with English and Korean sarcastic utterances," *Speech Commun.*, vol. 129, pp. 1–6, 2021. DOI: 10.1016/J.SPECOM.2021.02.002.
- [4] M. K. Luka, F. Ibikunle, and O. Gregory, "Neural network based Hausa language speech recognition," *Int. J. Adv. Res. Artif. Intell. Sci. Inf. Organ.*, vol. 1, no. 2, pp. 39–44, 2012. DOI: 10.14569/IJARAI.2012.010207.
- [5] G. Dharmale, V. M. Thakare, and D. D. Patil, "Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English," *Int. J. Adv. Comput. Sci. Appl. Sci. Inf. Organ.*, 2019. DOI: 10.14569/IJACSA.2019.0100212.
- [6] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," in *2016 6th international conference-cloud system and big data engineering (confluence)*, 2016, pp. 493–497. DOI: 10.1109/CONFLUENCE.2016.7508170.
- [7] S. Karpagavalli and E. Chandra, "A review on automatic speech recognition architecture and approaches," *Int. J. Signal Process. Image Process. Pattern Recognit.*, vol. 9, no. 4, pp. 393–404, 2016. DOI: 10.14257/ijisp.2016.9.4.34.
- [8] A. A. A. Zamil, S. Hasan, S. M. D. J. Baki, J. M. D. Adam, and I. Zaman, "Emotion detection from speech signals using voting mechanism on classified frames," in *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, 2019, pp. 281–285. DOI: 10.1109/ICREST.2019.8644168.
- [9] K. Jermstittiparsert *et al.*, "Pattern recognition and features selection for speech emotion recognition model using deep learning," *Int. J. Speech Technol.*, vol. 23, no. 4, pp. 799–806, 2020. DOI: 10.1007/s10772-020-09690-2.
- [10] A. Hussein, "Analysis of Voice Recognition Algorithms using MATLAB," *Int. J. Eng. Res. & Technol.*, vol. 4, no. 8, pp. 273–278, 2015.
- [11] W. Liu, Q. Liao, F. Qiao, W. Xia, C. Wang, and F. Lombardi, "Approximate designs for fast Fourier transform (FFT) with application to speech recognition," *IEEE Trans. Circuits Syst. I Regul. Pap.*, vol. 66, no. 12, pp. 4727–4739, 2019. DOI: 10.1109/TCSI.2019.2933321.
- [12] S. Pandita and S. K. Sharma, "Speech-enhancement techniques and review on the role of labview in speech-enhancement techniques," *INROADS-An Int. J. Jaipur Natl. Univ.*, vol. 8, no. 1and2, pp. 31–34, 2019. DOI: 10.5958/2277-4912.2019.00005.5.
- [13] F. Barkani, H. Satori, M. Hamidi, O. Zealouk, and N. Laaidi, "Comparative Evaluation of Speech Recognition Systems Based on Different Toolkits," *Embed. Syst. Artif. Intell.*, pp. 33–41, 2020. DOI: 10.1007/978-981-15-0947-6\_4.
- [14] N. Berezowski and M. Haid, "How to Use a Graphical Programming Language in Functional Safety, using the Example of Lab VIEW," in *2020 IEEE International RF and Microwave Conference (RFM)*, 2020, pp. 1–4. DOI: 10.1109/RFM50841.2020.9344729.
- [15] S. Pleshkova, Z. Zahariev, and A. Bekiarshi, "Development of Speech Recognition Algorithm and LabView Model for Voice Command Control of Mobile Robot Motion," in *2018 International Conference on High Technology for Sustainable Development (HiTech)*, 2018, pp. 1-5. DOI: 10.1109/HiTech.2018.8566257.
- [16] L. Keviczky, R. Bars, J. Hetthéssy, and C. Bányász, "Introduction to MATLAB," in *Control Engineering: MATLAB Exercises*, Springer, 2019, pp. 1–27.
- [17] I. McLoughlin. *Speech and Audio Processing: A MATLAB®-based Approach*. University Printing House, Cambridge CB2 8BS, United Kingdom, 2016.
- [18] M. Alzantot, B. Balaji, and M. Srivastava, "Did you hear that? adversarial examples against automatic speech recognition," *arXiv Prepr. arXiv1801.00554*, 2018.
- [19] B. H. Prasetyo and D. Syaury, "Implementation of Lossless Voice Data Communication using Network Streams on Embedded System," *JITECS (Journal Inf. Technol. Comput. Sci.)*, vol. 2, no. 2, 2017. DOI: 10.25126/jitecs.20172225.
- [20] A. Ghorbel, N. Ben Amor, and M. Jallouli, "Towards a Power Adaptation Strategy in Multi-core Embedded Devices. A Case Study: a HMI for Wheelchair Command Technique," *Ada User J.*, vol. 38, no. 2, 2017.
- [21] N. Hossain and M. Nazin, "Emovoice: Finding my mood from my voice signal," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 1826–1828. DOI: 10.1145/3267305.3277832.
- [22] A. S. S. Kyi and T. Shimamura, "Pre-Emphasis and Linear Prediction Error Filters for Quality Improvement of Bone-Conducted Speech," in *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*, 2020, pp. 481–484. DOI: 10.1109/WIECON-ECE52138.2020.9397998.
- [23] B. H. Prasetyo, H. Tamura, and K. Tanno, "Deep time-delay Markov network for prediction and modeling the stress and emotions state transition," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020. DOI: 10.1038/s41598-020-75155-w.
- [24] F.-N. Landini, "A Pronunciation Scoring System for Second Language Learners," *Univ. Buenos Aires Fac. Ciencias Exactas y Nat. Dep. Comput. Buenos Aires*, 2017.
- [25] D. Prabakaran and S. Sriuppili, "Speech Processing: MFCC Based Feature Extraction Techniques- An Investigation," *J. Physics*, vol. 1717, no. 1, pp. 1-8, 2021. DOI: 10.1088/1742-6596/1717/1/012009.
- [26] A. Sithara, A. Thomas, and D. Mathew, "Study of MFCC and IHC Feature Extraction Methods With Probabilistic Acoustic Models for Speaker Biometric Applications," *Proc. Comput. Sci.* vol. 143, pp. 267-276, 2018. DOI: 10.1016/j.procs.2018.10.395.
- [27] K. K. Sabor, A. Hamou-Lhadj, and A. Larsson, "Durfex: a feature extraction technique for efficient detection of duplicate bug reports," in

- 2017 IEEE international conference on software quality, reliability and security (QRS), 2017, pp. 240–250. DOI: 10.1109/QRS.2017.35.
- [28] J.-S. Bae, H. Bae, Y.-S. Joo, J. Lee, G.-H. Lee, and H.-Y. Cho, “Speaking Speed Control of End-to-End Speech Synthesis using Sentence-Level Conditioning,” *arXiv Prepr. arXiv2007.15281*, 2020.
- [29] B. H. Prasetyo, H. Tamura, and K. Tanno, “Emotional variability analysis based i-vector for speaker verification in under-stress conditions,” *Electronics*, vol. 9, no. 9, p. 1420, 2020. DOI: 10.3390/electronics9091420.
- [30] A. Labied, M.; Belangour, “Automatic Speech Recognition Features Extraction Techniques: A Multi-criteria Comparison,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 177–182, 2021. DOI: 10.14569/IJACSA.2021.0120821.
- [31] B. H. Prasetyo and D. Syauqy, “Design of Speaker Verification using Dynamic Time Warping (DTW) on Graphical Programming for Authentication Process,” *JITeCS (Journal Inf. Technol. Comput. Sci.)*, vol. 2, no. 1, pp. 11–18, 2017. DOI: 10.25126/jitecs.20172124.
- [32] G. Kurniawati and O. Karnalim, “Introducing a practical educational tool for correlating algorithm time complexity with real program execution,” *JITeCS (Journal Inf. Technol. Comput. Sci.)*, vol. 3, no. 1, pp. 1–15, 2018. DOI: 10.25126/jitecs.20183140.

# A New Priority Rule for Initial Ordering of Jobs in Permutation Flowshop Scheduling Problems

B. Dhanasakkaravarthi<sup>1</sup>

Research Scholar, School of Mechanical Engineering  
Sathyabama Institute of Science and Technology, India

A. Krishnamoorthy<sup>2</sup>

Professor, School of Mechanical Engineering  
Sathyabama Institute of Science and Technology, India

**Abstract**—Scheduling in a permutation flowshop refers to processing of jobs in a set of available machines in the same order. Among the several possible performance characteristics of a flowshop, makespan remains one of the highest preferred metrics by researchers in the past six decades. The constructive heuristic proposed by Nawaz-Enscore-Ham (NEH) is one of the best for makespan minimization. The performance essentially depends on the initial ordering jobs according to a particular priority rule (PR). The popular priority rules are non-increasing order of the jobs' total processing time, the sum of average processing time and standard deviation and, the sum of average processing time, standard deviation and absolute skewness among others. The objective of this paper is to propose and analyse a new job priority rule for the permutation flowshop. The popular priority rules available in the literature are studied and, one of the best priority rules; the sum of average processing time and standard deviation is slightly modified, by replacing the standard deviation by mean absolute deviation (MAD). To assess the performance of the new rule, four benchmark datasets are used. The computational results and statistical analyses demonstrate the better performance of the new rule.

**Keywords**—Priority rule; flowshop scheduling; makespan; NEH algorithm

## I. INTRODUCTION

Permutation flowshop scheduling problems (PFSSP) remain one of the most studied domains in operations research in the past six decades. PFSSP refers to scheduling ' $n$ ' number of jobs for processing in ' $m$ ' number of machines in the same order. According to Rinnooy Kan [1], PFSSP is proved to be NP-hard when the number of machines is greater than three. Since the PFSSP is np-hard, the computation time grows exponentially for larger problems and hence, exact solution becomes impossible or expensive. As a result, several dispatching rules, heuristics and metaheuristics have been proposed over the decades. Efficient heuristics report solutions with acceptable accuracy levels in a reasonable time. Priority rules or dispatching rules are some form of heuristics and have been studied in both academia and industry domains extensively for decades (Tay and Ho 2008)[2]. A few simple dispatching rules that are being extensively used in jobs' scheduling are: Shortest Processing Time (SPT) rule, Longest Processing Time (LPT) rule, Earliest Due Date (EDD) rule, First Come First Serve (FCFS) rule.

The optimization parameters for the PFSSP are generally; flow time, idle time or makespan (Liu et al., 2016) [3] to satisfy different production line requirements. They find

numerous real-time applications and could be combined with Internet of Things (IOT) for specific applications (Salis, 2021) [4]. Among the many parameters that are being optimised, makespan minimization is widely considered by researchers over the years. Johnson's [5] algorithm proposed in 1954 yields an optimum solution for two machines and ' $n$ ' jobs PFSSPs which was extended to three machines cases.

As the problem is NP-hard, the exact solution becomes impossible for larger problems and the computation time grows exponentially with the problem size. Earlier approximate heuristics could not yield the expected accuracy and the breakthrough came in 1983 when Nawaz-Enscore-Ham (NEH) [6] algorithm was proposed. NEH algorithm which has a complexity of  $O(n^3.m)$  uses the largest processing time (LPT) dispatching rule and is considered as one of the best constructive heuristics for makespan minimization even today. Many improvements and extensions have been proposed by many authors over the years. NEH essentially consists of three phases:

- Pre-arranging the jobs according to the non-increasing order of their total processing times (Priority Rule).
- Selecting the first two jobs from the processed sequence as the initial partial sequence (Initial Sequence).
- Inserting other jobs one by one at a suitable place that minimises the partial makespan (Insertion Phase).

Ribas et al. (2010) [7] tested four priority rules combined with the powerful insertion technique of NEH including rules from NEHKK1 (Kalczyński and Kamburowski 2008) [8], N&M (Nagano and Moccellini 2002) [9], LPT, and a random job sequence. Baskar and Xavier (2015) [10] analysed the job insertion technique for different initial sequences.

The authors improved the solution quality of NEH by suitably modifying the first and third phases. Several priority rules for NEH are proposed and available in the literature and the ones proposed by Dong et al. [11] and Liu et al. [12] yield better results.

Framinan et al. [13] analysed 177 initial sequences for the NEH heuristic and concluded that SUM PIJ / DECR (i.e. original NEH) is ranked 1 among all for the makespan minimisation.

During the insertion phase, we come across several occasions when the partial makespan remains the same for

more than one partial sequence. NEH breaks such ties randomly. Several effective tie-breaking rules are proposed and analysed for the solution quality. The tie-breaking rules proposed by Fernandez and Framinan [14], Lie et al. [12], Benavides [15] and the ones recently by Baskar and Xavior [16] are a few to mention.

This paper considers the sum of the average processing times and means absolute deviation (MAD) for the initial ordering of jobs (Priority Rule) and analyses its impact using well-known benchmark datasets.

The structure of this paper is as follows: the new priority rule is presented in Section 2 followed by the benchmark and performance metrics used for the assessment in Section 3. The results and statistical analyses are detailed in Section 4 and Section 5 discusses about the conclusion, limitations and future work.

## II. NEW PRIORITY RULE

The initial ordering of jobs or priority rule does affect the solution quality of the NEH heuristic. The priority rule considered by the original NEH is non-increasing order of the jobs' total processing times. Total processing time (TPT) for a particular job,  $j$  can be represented as,

$$TPT_j = \sum_{i=1}^m p_{ij} \quad (1)$$

$m$  – Number of machines in the schedule.  $p_{ij}$  denotes the processing time of job ' $j$ ' in the machine ' $i$ '.

Dong et al. [11] added the standard deviation of the processing times with the average total processing time,  $AVG_j$  for a job and reported improved results. Mathematically, standard deviation,

$$SD_j = \sqrt{\frac{1}{(m-1)} \sum_{i=1}^m (p_{ij} - AVG_j)^2} \quad (2)$$

Another priority rule proposed recently by Liu et al. [12] adds the absolute skewness with the priority rule of Dong et al. [11] and the rule improves the solution further. Skewness (SKE) for a job,  $j$  is defined as,

$$SKE_j = \frac{\frac{1}{m} \sum_{i=1}^m (p_{ij} - AVG_j)^3}{\left[ \sqrt{\frac{1}{(m-1)} \sum_{i=1}^m (p_{ij} - AVG_j)^2} \right]^3} \quad (3)$$

In this paper, the standard deviation is replaced by another similar metric, the Mean Absolute Deviation (MAD) which is the average of the absolute deviations from the mean value.

For a job ' $j$ ',

$$MAD_j = \frac{1}{m} \sum_{i=1}^m |(p_{ij} - AVG_j)| \quad (4)$$

MAD was used as the variation measure in the tie-breaking rule of Liu et al. [12]. Another metric, the median absolute deviation (MAD1) is also used instead of mean absolute deviation during the process; that is, descending order of average processing time + median absolute deviation as a priority rule. However, the results are not encouraging in this case.

For a job ' $j$ ',

$$MAD1_j = \text{Median} (|(p_{ij} - AVG_j)|) \quad (5)$$

The benchmarks and analyses results using different combinations are presented and discussed in the coming sections.

## III. BENCHMARK AND PERFORMANCE MATRIC USED

For comparing the performance of similar heuristic algorithms in flowshop scheduling, several benchmark problem sets are available in the literature. One of the earlier benchmarks proposed for permutation flowshop is the one proposed by Carlier in 1978 [17] with varying jobs and machines combination. The instances have high processing times, up to 999-time units. DUM dataset proposed in 1998 by Demirkol et.al. [18] has a combination of jobs {20, 30, 40, 50} and machines {15, 20} resulting in 40 instances of 5 in each set. The processing time is randomly fixed between {0, 200} time units. The Taillard [19] instances proposed in 1993 are available under 12 groups of 10 instances each. The processing times in 5, 10 and 20 machines vary from 0 to 99 for 20, 50, 100, 200 and 500 numbers of jobs. The known upper bounds for all the 120 problems are available online [20] and are accessible for researchers.

Recently, Vallada et al. [21] proposed new hard benchmarks which are 480 in numbers. They are categorized into small and large problem sets. The small instances, termed as VFR (Small) are 240 in numbers with a combination of {10, 20, 30, 40, 50, 60} jobs and {5, 10, 15, 20} machines. Similarly, the larger instances have a combination of {100, 200, 300, 400, 500, 600, 700, 800} jobs and {20, 40, 60} machines, totaling 240 in numbers. The instances are grouped under 48 sets of 10 instances each. In line with Taillard's, these benchmarks do have processing times from 0 to 99. The known upper bounds are provided by the authors themselves.

The parameter used for the comparison of the performance of different heuristics is the Percentage Relative Deviation (PRD) which is defined as:

$$\text{Percentage Relative Deviation, PRD} = \frac{(\text{Makespan} - \text{KnownUpperBound}) \times 100}{\text{KnownUpperBound}} \quad (6)$$

## IV. COMPUTATIONAL RESULTS AND DISCUSSION

Different priority rules (initial ordering of jobs) considered are summarized below:

- PR-NEH: Descending order of their total processing times.
- PR-D: Descending order of the sum of their average processing time and standard deviation.
- PR-LJP: Descending order of the sum of their average processing time, standard deviation and absolute (skewness).
- PR-DB: Descending order of the sum of their average processing time and mean absolute deviation.

- “No PR” refers to no initial ordering of jobs; raw data are used as received.

Codes are generated for all priority rules in MATLAB R2012b and run in an i5 desktop PC with 4 GB RAM. Many heuristics initially order the jobs according to a priority rule and then refine the solution further using some strategy. The initial priority ordering may have significant impact on the final result. To start with, Taillard instances are initially ordered as per the priority rule (PR) considered and the makespans computed without any further processing. The average PRDs (APRDs) are presented in Table I. The bold digits indicate the best result for a particular problem set. There is not much difference in the results among the PRs considered. The best APRD of 21.40% is reported by PR-DB and the worst one by PR-D (21.64%); the difference being 1.12%. Even no “no priority rule” reports a better mean value of 21.61% and accounts for best results in 4/12 sets. Though PR-DB reports the lowest deviation, it reports best results in one case only, 10 machines and 100 jobs.

Subsequently, the impact of different initial sequences has been investigated. NEH considers the first two jobs of the processed sequence as the initial sequence. Baskar et al. [22] considered a few other sets of jobs as initial sequences and analysed (Table II) the impact using Taillard dataset. The performance was analysed by randomly selecting two jobs also from the processed sequence. The results show that the APRDs vary slightly with respect to the initial partial sequence. According to them, randomly selecting two jobs also results in a reasonably good APRD with 3.43% which may be due to the job insertion strategy that was originally used by the NEH algorithm. It is to be noted that many heuristics could not even better this mean value of 3.43% for the Taillard benchmark.

Now, to assess the performance; the PR is applied to the classic NEH algorithm and the results are tabulated in Table III to Table VI. The benchmark used are; Carlier’s [17] proposed in 1978, Taillard’s [19], in 1993, Demirkol’s [18] in 1998 and the latest VFR benchmark proposed in 2015 by Vallada et al.[21]. The better APRDs are shown in bold and italics in all the tables.

For the Carlier 8 instances (Table III); Dong et al. priority rule, PR-D reports an APRD of 1.44%, an increase of 3.60% in the APRD over PR-NEH which is taken as the reference for the analysis in this paper. That is, for this dataset, the solution quality deteriorates when PR-D is used for initial ordering the jobs. The APRD of the new priority rule, PR-DB is 1.21%, a significant 12.95% improvement over the rule used by NEH.

For the comparatively larger dataset of Demirkol’s which are 40 in numbers also, the proposed priority rule, PR-DB reports better APRD than the priority rules, PR-NEH and PR-D with an improvement of 4.70% over the reference PR-NEH. The improvement of the rule, PR-D over PR-NEH being 1.61%. Here, the lower bounds (LB) are considered instead of upper bounds (UB) for better comparison as the makespans better the UBs provided by Demirkol et al. in many cases. While using Taillard’s dataset; three more priority rules are considered as detailed below:

- PR-LJP1: Descending order of the sum of their average processing time, mean absolute deviation and absolute (skewness).
- PR-LJP2: Descending order of the sum of their average processing time, mean absolute deviation and skewness.
- PR-DB1: Descending order of the sum of their average processing time and median absolute deviation.

TABLE I. APRDS OF PRIORITY RULES – TAILLARD BENCHMARK

| Size (mxj) | No PR        | PR-NEH       | PR-D         | PR-LJP       | PR-DB        |
|------------|--------------|--------------|--------------|--------------|--------------|
| 5x20       | <b>24.98</b> | 26.27        | 27.41        | 27.66        | 26.24        |
| 10x20      | 28.77        | 28.47        | 28.15        | <b>27.96</b> | 28.16        |
| 20x20      | 21.43        | 21.50        | 20.81        | <b>20.56</b> | 21.31        |
| 5x50       | <b>15.32</b> | 16.46        | 17.21        | 17.63        | 16.91        |
| 10x50      | <b>25.05</b> | 28.86        | 27.10        | 26.66        | 26.34        |
| 20x50      | 29.73        | 29.15        | 28.13        | <b>27.76</b> | 27.89        |
| 5x100      | 13.63        | 12.06        | 12.22        | <b>11.98</b> | 12.72        |
| 10x100     | 20.92        | 19.08        | 20.21        | 20.24        | <b>18.85</b> |
| 20x100     | 25.51        | <b>23.50</b> | 24.47        | 24.77        | 24.61        |
| 10x200     | 15.67        | 15.70        | <b>14.78</b> | 15.31        | 15.11        |
| 20x200     | 22.28        | <b>21.39</b> | 22.32        | 21.74        | 22.30        |
| 20x500     | <b>15.99</b> | 16.31        | 16.80        | 16.51        | 16.33        |
| Mean       | 21.61        | 21.56        | 21.64        | 21.56        | 21.40        |

TABLE II. APRDS OF INITIAL PARTIAL SEQUENCES APPLIED TO NEH - TAILLARD BENCHMARK

| Size (mxj) | Jobs 1 and 2 (NEH) | Middle 2 Jobs | Jobs 1 and 3 | Jobs 1 and 4 | Randomly 2 Jobs |
|------------|--------------------|---------------|--------------|--------------|-----------------|
| 5x20       | 3.30               | 2.79          | 3.03         | 3.10         | 3.89            |
| 10x20      | 4.60               | 3.68          | 5.04         | 4.18         | 4.40            |
| 20x20      | 3.73               | 3.67          | 3.76         | 3.58         | 3.79            |
| 5x50       | 0.73               | 0.82          | 0.68         | 0.68         | 0.94            |
| 10x50      | 5.07               | 5.36          | 4.90         | 4.78         | 5.39            |
| 20x50      | 6.65               | 6.54          | 6.63         | 6.66         | 6.85            |
| 5x100      | 0.53               | 0.51          | 0.51         | 0.50         | 0.56            |
| 10x100     | 2.21               | 2.11          | 2.20         | 2.19         | 2.24            |
| 20x100     | 5.34               | 5.72          | 5.19         | 5.47         | 5.34            |
| 10x200     | 1.26               | 1.41          | 1.24         | 1.30         | 1.35            |
| 20x200     | 4.41               | 4.07          | 4.55         | 4.39         | 4.35            |
| 20x500     | 2.07               | 2.26          | 2.12         | 2.07         | 2.12            |
| Mean       | 3.32               | 3.24          | 3.32         | 3.24         | 3.43            |

TABLE III. APRDS OF PRIORITY RULES APPLIED TO NEH- CARLIER BENCHMARK

| Size (mxj)   | BM    | UB   | PR-NEH | PRD  | PR-D | PRD   | PR-DB | PRD   |
|--------------|-------|------|--------|------|------|-------|-------|-------|
| 5x11         | Carl1 | 7038 | 7038   | 0    | 7038 | 0     | 7038  | 0     |
| 4x13         | Carl2 | 7166 | 7376   | 2.93 | 7376 | 2.93  | 7376  | 2.93  |
| 5x12         | Carl3 | 7312 | 7399   | 1.19 | 7399 | 1.19  | 7399  | 1.19  |
| 4x14         | Carl4 | 8003 | 8003   | 0    | 8129 | 1.57  | 8021  | 0.22  |
| 6x10         | Carl5 | 7720 | 7835   | 1.49 | 7843 | 1.59  | 7843  | 1.59  |
| 9x8          | Carl6 | 8505 | 8773   | 3.15 | 8773 | 3.15  | 8570  | 0.76  |
| 7x7          | Carl7 | 6590 | 6590   | 0    | 6590 | 0     | 6590  | 0     |
| 8x8          | Carl8 | 8366 | 8564   | 2.37 | 8457 | 1.09  | 8617  | 3.00  |
| Mean         |       |      |        | 1.39 |      | 1.44  |       | 1.21  |
| %Improvement |       |      |        | Ref. |      | -3.60 |       | 12.95 |

TABLE IV. APRDS (FROM LB) OF PRIORITY RULES APPLIED TO NEH- DEMIRKOL BENCHMARK

| Instance        | LB   | PR-NEH | PRD   | PR-D | PRD   | PR-DB | PRD   |
|-----------------|------|--------|-------|------|-------|-------|-------|
| flcmax_20_15_3  | 3354 | 4071   | 21.38 | 4018 | 19.80 | 4065  | 21.20 |
| flcmax_20_15_6  | 3168 | 3898   | 23.04 | 3878 | 22.41 | 3870  | 22.16 |
| flcmax_20_15_4  | 2997 | 3672   | 22.52 | 3617 | 20.69 | 3625  | 20.95 |
| flcmax_20_15_10 | 3420 | 4248   | 24.21 | 4223 | 23.48 | 4217  | 23.30 |
| flcmax_20_15_5  | 3494 | 4007   | 14.68 | 4028 | 15.28 | 4038  | 15.57 |
| flcmax_20_20_1  | 3776 | 4779   | 26.56 | 4641 | 22.91 | 4674  | 23.78 |
| flcmax_20_20_3  | 3758 | 4567   | 21.53 | 4535 | 20.68 | 4598  | 22.35 |
| flcmax_20_20_9  | 3902 | 4699   | 20.43 | 4666 | 19.58 | 4611  | 18.17 |
| flcmax_20_20_2  | 3881 | 4606   | 18.68 | 4619 | 19.02 | 4626  | 19.20 |
| flcmax_20_20_10 | 3823 | 4487   | 17.37 | 4515 | 18.10 | 4462  | 16.71 |
| flcmax_30_15_3  | 4020 | 4770   | 18.66 | 4729 | 17.64 | 4692  | 16.72 |
| flcmax_30_15_4  | 4080 | 4912   | 20.39 | 4924 | 20.69 | 4890  | 19.85 |
| flcmax_30_15_9  | 4022 | 4857   | 20.76 | 4737 | 17.78 | 4770  | 18.60 |
| flcmax_30_15_8  | 4490 | 5070   | 12.92 | 5056 | 12.61 | 4982  | 10.96 |

|                 |      |      |              |      |              |      |              |
|-----------------|------|------|--------------|------|--------------|------|--------------|
| flcmax_30_15_6  | 4184 | 5041 | 20.48        | 5013 | 19.81        | 4888 | <b>16.83</b> |
| flcmax_30_20_3  | 4806 | 5664 | 17.85        | 5648 | 17.52        | 5561 | <b>15.71</b> |
| flcmax_30_20_1  | 4772 | 5891 | <b>23.45</b> | 5995 | 25.63        | 5940 | 24.48        |
| flcmax_30_20_6  | 5004 | 5919 | <b>18.29</b> | 5989 | 19.68        | 5970 | 19.30        |
| flcmax_30_20_10 | 4899 | 5523 | 12.74        | 5532 | 12.92        | 5464 | <b>11.53</b> |
| flcmax_30_20_2  | 4757 | 5629 | 18.33        | 5470 | <b>14.99</b> | 5591 | 17.53        |
| flcmax_40_15_5  | 5560 | 6286 | 13.06        | 6380 | 14.75        | 6193 | <b>11.38</b> |
| flcmax_40_15_9  | 5119 | 5931 | 15.86        | 5907 | <b>15.39</b> | 5947 | 16.18        |
| flcmax_40_15_2  | 5290 | 6113 | 15.56        | 6105 | 15.41        | 6102 | <b>15.35</b> |
| flcmax_40_15_10 | 5596 | 6206 | 10.90        | 6271 | 12.06        | 6115 | <b>9.27</b>  |
| flcmax_40_15_8  | 5576 | 6394 | 14.67        | 6329 | <b>13.50</b> | 6347 | 13.83        |
| flcmax_40_20_3  | 5693 | 6816 | <b>19.73</b> | 6865 | 20.59        | 6866 | 20.60        |
| flcmax_40_20_9  | 5998 | 6929 | <b>15.52</b> | 7065 | 17.79        | 6995 | 16.62        |
| flcmax_40_20_6  | 5990 | 7154 | 19.43        | 7160 | 19.53        | 7097 | <b>18.48</b> |
| flcmax_40_20_7  | 6170 | 7026 | <b>13.87</b> | 7107 | 15.19        | 7080 | 14.75        |
| flcmax_40_20_5  | 6011 | 6910 | 14.96        | 6846 | 13.89        | 6842 | <b>13.82</b> |
| flcmax_50_15_6  | 6290 | 7264 | 15.48        | 7206 | 14.56        | 7111 | <b>13.05</b> |
| flcmax_50_15_5  | 6355 | 6928 | <b>9.02</b>  | 7026 | 10.56        | 6972 | 9.71         |
| flcmax_50_15_1  | 6198 | 6909 | 11.47        | 6860 | <b>10.68</b> | 6916 | 11.58        |
| flcmax_50_15_8  | 6312 | 7180 | 13.75        | 7158 | 13.40        | 7132 | <b>12.99</b> |
| flcmax_50_15_2  | 6531 | 7330 | 12.23        | 7267 | <b>11.27</b> | 7278 | 11.44        |
| flcmax_50_20_2  | 6740 | 8138 | 20.74        | 8063 | 19.63        | 8021 | <b>19.01</b> |
| flcmax_50_20_1  | 6736 | 7602 | 12.86        | 7550 | <b>12.08</b> | 7673 | 13.91        |
| flcmax_50_20_7  | 6756 | 7965 | <b>17.90</b> | 8081 | 19.61        | 7993 | 18.31        |
| flcmax_50_20_8  | 6897 | 7924 | 14.89        | 7890 | 14.40        | 7617 | <b>10.44</b> |
| flcmax_50_20_4  | 6830 | 8256 | 20.88        | 8218 | 20.32        | 8098 | <b>18.57</b> |
| <b>Mean</b>     |      |      | <b>17.43</b> |      | <b>17.15</b> |      | <b>16.61</b> |
| %Improvement    |      |      | Ref.         |      | 1.61         |      | 4.70         |

TABLE V. APRDS OF PRIORITY RULES APPLIED TO NEH- TAILLARD BENCHMARK

| Size (mxj)   | PR-NEH | PR-D        | PR-LJP      | PR-LJP1     | PR-LJP2     | PR-DB       | PR-DB1      |
|--------------|--------|-------------|-------------|-------------|-------------|-------------|-------------|
| 5x20         | 3.30   | 2.70        | 2.71        | 2.95        | 3.04        | 2.74        | <b>1.93</b> |
| 10x20        | 4.60   | 4.08        | 3.68        | 4.02        | <b>3.37</b> | 3.87        | 4.89        |
| 20x20        | 3.73   | 3.82        | <b>2.91</b> | 3.20        | 3.13        | 2.95        | 3.73        |
| 5x50         | 0.73   | 0.89        | 0.88        | <b>0.70</b> | 0.80        | 0.92        | 0.93        |
| 10x50        | 5.07   | 4.90        | 4.84        | <b>4.71</b> | 4.82        | 5.44        | 5.56        |
| 20x50        | 6.65   | 6.12        | 6.42        | 6.50        | 6.67        | 6.45        | <b>6.08</b> |
| 5x100        | 0.53   | <b>0.41</b> | 0.54        | 0.57        | 0.50        | 0.50        | 0.60        |
| 10x100       | 2.21   | <b>2.16</b> | 2.24        | 2.33        | 2.36        | 2.40        | 2.39        |
| 20x100       | 5.34   | 5.65        | <b>4.99</b> | 5.28        | 5.07        | 5.16        | 5.71        |
| 10x200       | 1.26   | 1.27        | 1.24        | <b>1.23</b> | 1.31        | 1.29        | <b>1.23</b> |
| 20x200       | 4.41   | 4.57        | <b>4.14</b> | 4.24        | 4.26        | 4.27        | 4.39        |
| 20x500       | 2.07   | 2.12        | 2.12        | 2.14        | 2.10        | <b>2.06</b> | 2.09        |
| Mean         | 3.32   | 3.22        | 3.06        | 3.16        | 3.12        | 3.17        | 3.29        |
| %Improvement | Ref.   | 3.01        | 7.83        | 4.82        | 6.02        | 4.52        | 0.91        |

TABLE VI. AVERAGE PERCENT RELATIVE DEVIATION OF DIFFERENT PRIORITY RULES FOR VFR BENCHMARK

| Size (mxj)   | PR-NEH      | PR-D        | PR-DB       | Size (mxj)   | PR-NEH      | PR-D        | PR-DB       |
|--------------|-------------|-------------|-------------|--------------|-------------|-------------|-------------|
| 5x10         | 2.18        | <b>1.51</b> | 1.67        | 20x100       | 5.71        | 5.61        | <b>5.27</b> |
| 10x10        | 1.63        | <b>1.46</b> | 2.66        | 40x100       | 5.67        | <b>5.31</b> | 5.43        |
| 15x10        | <b>1.53</b> | 2.17        | 2.47        | 60x100       | 4.95        | <b>4.51</b> | 4.85        |
| 20x10        | 1.99        | <b>1.52</b> | 1.59        | 20x200       | 4.23        | <b>4.04</b> | 4.19        |
| 5x20         | <b>1.51</b> | 2.76        | 2.19        | 40x200       | 4.71        | 4.66        | <b>4.51</b> |
| 10x20        | <b>4.82</b> | 4.93        | 5.34        | 60x200       | 4.55        | 4.35        | <b>4.28</b> |
| 15x20        | 4.33        | <b>3.93</b> | 4.50        | 20x300       | 3.00        | 3.03        | <b>2.99</b> |
| 20x20        | 4.12        | <b>3.50</b> | 3.89        | 40x300       | 4.08        | 3.90        | <b>3.88</b> |
| 5x30         | <b>1.43</b> | 1.64        | 1.45        | 60x300       | 3.93        | <b>3.91</b> | 3.93        |
| 10x30        | <b>5.26</b> | 5.46        | 5.29        | 20x400       | 2.58        | 2.46        | <b>2.24</b> |
| 15x30        | 5.83        | 5.44        | <b>5.31</b> | 40x400       | 3.66        | 3.51        | <b>3.43</b> |
| 20x30        | <b>5.41</b> | 5.49        | 5.47        | 60x400       | 3.56        | 3.47        | <b>3.42</b> |
| 5x40         | 1.09        | 0.79        | <b>0.70</b> | 21x500       | 2.27        | 2.23        | <b>2.00</b> |
| 10x40        | 4.97        | <b>4.52</b> | 4.75        | 40x500       | 3.20        | 3.11        | <b>3.06</b> |
| 15x40        | 6.05        | 5.87        | <b>5.56</b> | 60x500       | <b>3.12</b> | 3.20        | 3.13        |
| 20x40        | <b>5.14</b> | 5.29        | 5.58        | 20x600       | <b>1.57</b> | 1.64        | 1.62        |
| 5x50         | <b>0.55</b> | 0.82        | 0.84        | 40x600       | 3.13        | 2.98        | <b>2.86</b> |
| 10x50        | 4.58        | <b>4.45</b> | 4.59        | 60x600       | 2.93        | 2.94        | <b>2.92</b> |
| 15x50        | <b>6.52</b> | 6.90        | 6.76        | 20x700       | 1.40        | <b>1.23</b> | 1.32        |
| 20x50        | <b>5.96</b> | 6.00        | 6.71        | 40x700       | 2.77        | <b>2.60</b> | 2.70        |
| 5x60         | 0.89        | 0.48        | <b>0.34</b> | 60x700       | 2.75        | <b>2.71</b> | 2.78        |
| 10x60        | 3.96        | <b>3.94</b> | 4.25        | 20x800       | 1.23        | 1.15        | <b>1.14</b> |
| 15x60        | <b>5.79</b> | 5.91        | 5.91        | 40x800       | <b>2.43</b> | 2.52        | 2.50        |
| 20x60        | 6.45        | 6.42        | <b>6.14</b> | 60x800       | 2.71        | <b>2.67</b> | 2.68        |
| Mean         | 3.83        | 3.80        | 3.92        | Mean         | 3.34        | 3.24        | 3.21        |
| %Improvement | Ref.        | 0.78        | -2.35       | %Improvement | Ref.        | 2.99        | 3.89        |

The PR-LJP proposed by Liu et al. [12] is proved to yield better results for the Taillard dataset. PR-LJP adds a third metric, absolute (skewness) to the priority rule, PR-D. However, when this absolute (skewness) is added to the sum of average processing time and mean absolute deviation (PR-LJP1); the APRD comes down to 3.16% from 3.06%. The APRD slightly improves to 3.12% when the absolute (skewness) is replaced by skewness (PR-LJP2). For this dataset, PR-D reports an APRD of 3.22% and PR-DB, 3.17%. The improvement of PR-DB is 4.52% over PR-NEH which is better than PR-D (3.01% improvement). When the median absolute deviation (MAD1) replaces the mean absolute deviation (PR-DB1), we get an APRD of 3.29% (Table V) for the Taillard instances.

The results of VFR (Small) and VFR (Large) datasets are given in Table VI. The performance is different here. For the VFR (Small) dataset, the performance of PR-DB worsens by 2.35% whereas; PR-D reports 0.78% improvement over the reference priority rule, PR-NEH. For the smaller problems, the priority rule, PR-D reports the best results. However, when the number of problem sets are considered, the reference rule, PR-NEH accounts for better results in 11/24 sets followed by PR-

D, 8/24 sets and the new rule, PR-DB does well in only 5/24 cases.

However, for the VFR (Large) dataset, the performance of the newly proposed rule, PR-DB is better than PR-D and PR-NEH. The improvement over PR-NEH is 3.89% for this larger dataset. The new rule also reports best results in 13/24 problem sets. PR-D comes next with better results in 8/24 cases. The priority rule, PR-NEH which performs extremely well in smaller problems of VFR, could not match that performance and accounts for only 3/24 problem sets.

The summary of the results is presented in Table VII.

TABLE VII. SUMMARY OF RESULTS (APRDs)

| Benchmark          | PR-NEH | PR-D        | PR-DB        |
|--------------------|--------|-------------|--------------|
| Carrier            | 1.39   | 1.44        | <b>1.21</b>  |
| Demrikol (over LB) | 17.43  | 17.15       | <b>16.61</b> |
| Taillard           | 3.32   | 3.22        | <b>3.17</b>  |
| VFR (Small)        | 3.83   | <b>3.80</b> | 3.92         |
| VFR (Large)        | 3.34   | 3.24        | <b>3.21</b>  |

For all the datasets except VFR (Small), the newly proposed priority rule, PR-DB outperforms PR-D and PR-NEH.

Paired t-tests are carried out at 95% confidence level using MINITAB17 and the results are presented in Table VIII.

TABLE VIII. PAIRED T-TEST ON DIFFERENT BENCHMARKS

| Pairs              | T-Value | P-Value      |
|--------------------|---------|--------------|
| <b>Carlier</b>     |         |              |
| PR-NEH vs PR-D     | -0.18   | 0.862        |
| PR-NEH vs PR-DB    | 0.55    | 0.597        |
| <b>Demirkol</b>    |         |              |
| PR-NEH vs PR-D     | 1.30    | 0.202        |
| PR-NEH vs PR-DB    | 3.75    | <b>0.001</b> |
| <b>Taillard</b>    |         |              |
| PR-NEH vs PR-D     | 1.16    | 0.270        |
| PR-NEH vs PR-DB    | 1.46    | 0.172        |
| <b>VFR (Small)</b> |         |              |
| PR-NEH vs PR-D     | 0.38    | 0.710        |
| PR-NEH vs PR-DB    | -0.85   | 0.403        |
| <b>VFR (Large)</b> |         |              |
| PR-NEH vs PR-D     | 3.86    | <b>0.001</b> |
| PR-NEH vs PR-DB    | 4.54    | <b>0.000</b> |

Since only PR-D and PR-DB are similar priority rules, they are compared with PR-NEH and listed in Table VII and Table VIII. The analyses could prove the statistical significance in three cases with probability values of 0.001, 0.001 and 0.000 for the pairs; PR-NEH vs PR-DB (Demirkol), PR-NEH vs PR-D (VFR-Large) and PR-NEH vs PR-DB (VFR-Large) respectively.

## V. CONCLUSION AND FUTURE WORK

This paper proposes a new priority rule, PR-DB for the permutation flowshop scheduling problems. The popular NEH algorithm considers the jobs according to the non-increasing order of their total processing times. Dong et al. added another metric, the standard deviation of the processing times of a particular job for initial ordering. In this work, the standard deviation used in the priority rule of Dong et al., PR-D is replaced by Mean Absolute Deviation. The PR is applied to the classic NEH heuristic for makespan minimisation. Computational results show that the performance is better than PR-D for the Carlier, Demirkol, Taillard and VFR (Large) datasets. The results do not improve in the case of VFR (Small) instances. The paired t-tests are carried out to assess the statistical significance. The main advantage of the new rule is that it is simple yet powerful. It can be easily applied in scheduling the jobs in any engineering industry. Similarly, it can be combined with any popular tie-breaking rule to improve the solution quality further. Also, the results obtained could be used as the seed solution for any metaheuristic for better solution and schedules. Future work includes the assessment of PR-DB for other potential benchmarks also.

## ACKNOWLEDGMENT

The authors are sincerely thankful to the anonymous referees and editors, who provide constructive comments to improve the technical content and presentation of the paper. We also thank Professors Michele Lanzetta, Andrea Rossi and Reha Uzsoy for providing Demirkol dataset and other required information in writing this paper.

## REFERENCES

- [1] Rinnooy kan ah. Machine scheduling problems: classification, complexity, and computations. PhDthesis, University of Amsterdam. 1976.
- [2] Tay JC, Ho NB. Evolving dispatching rules using genetic programming for solving multi-objective flexible job-shop problems. Computers & Industrial Engineering. 2008 Apr 1;54(3):453-73.
- [3] Liu W, Jin Y, Price M. A new Nawaz–Enscore–Ham-based heuristic for permutation flow-shop problems with bicriteria of makespan and machine idle time. Engineering Optimization. 2016 Oct 2;48(10):1808-22.
- [4] Salis A. Towards the Internet of Behaviors in Smart Cities through a Fog-To-Cloud Approach. HighTech and Innovation Journal. 2021 Dec 1;2(4):273-84.
- [5] Johnson SM. Optimal two-and three-stage production schedules with setup times included. Naval research logistics quarterly. 1954 Mar;1(1):61-8.
- [6] Nawaz M, Enscore Jr EE, Ham I. A heuristic algorithm for the m-machine, n-job flow-shop sequencing problem. Omega. 1983 Jan 1;11(1):91-5.
- [7] Ribas I, Companys R, Tort-Martorell X. Comparing three-step heuristics for the permutation flowshop problem. Computers & Operations Research. 2010 Dec 1;37(12):2062-70.
- [8] Kalczynski PJ, Kamburowski J. An improved NEH heuristic to minimize makespan in permutation flow shops. Computers & Operations Research. 2008 Sep 1;35(9):3001-8.
- [9] Nagano MS, Moccellini JV. A high quality solution constructive heuristic for flow shop sequencing. Journal of the Operational Research Society. 2002 Dec;53(12):1374-9.
- [10] Baskar A, Xavier MA. Analysis of job insertion technique for different initial sequences in permutation flow shop scheduling problems. International Journal of Enterprise Network Management. 2015;6(3):153-74.
- [11] Dong X, Huang H, Chen P. An improved NEH-based heuristic for the permutation flowshop problem. Computers & Operations Research. 2008 Dec 1;35(12):3962-8.
- [12] Liu W, Jin Y, Price M. A new improved NEH heuristic for permutation flowshop scheduling problems. International Journal of Production Economics. 2017 Nov 1;193:21-30.
- [13] Framinan JM, Leisten R, Rajendran C. Different initial sequences for the heuristic of Nawaz, Enscore and Ham to minimize makespan, idletime or flowtime in the static permutation flowshop sequencing problem. International Journal of Production Research. 2003 Jan 1;41(1):121-48.
- [14] Fernandez-Viagas V, Framinan JM. On insertion tie-breaking rules in heuristics for the permutation flowshop scheduling problem. Computers & Operations Research. 2014 May 1;45:60-7.
- [15] Benavides AJ. A New Tiebreaker in the NEH heuristic for the Permutation Flow Shop Scheduling Problem. EasyChair; 2018 Sep 14.
- [16] Baskar A, Xavier MA. New idle time-based tie-breaking rules in heuristics for the permutation flowshop scheduling problems. Computers & Operations Research. 2021 Sep 1;133:105348.
- [17] Carlier J. Ordonnements a contraintes disjonctives. RAIRO-Operations Research. 1978;12(4):333-50.
- [18] Demirkol E, Mehta S, Uzsoy R. Benchmarks for shop scheduling problems. European Journal of Operational Research. 1998 Aug 16;109(1):137-41.
- [19] Taillard E. Benchmarks for basic scheduling problems. european journal of operational research. 1993 Jan 22;64(2):278-85.

- [20] Taillard E. Summary of best known lower and upper bounds of Taillard's instances. Available in <http://ina2.eivd.ch/collaborateurs/etd/problemes.dir/ordonnancement.dir/ordonnancement.html>. 2005.
- [21] Vallada E, Ruiz R, Framinan JM. New hard benchmark for flowshop scheduling problems minimising makespan. *European Journal of Operational Research*. 2015 Feb 1;240(3):666-77.
- [22] Baskar A, Xavier MA, Dhanasakkaravarthi B. Impact of Initial Partial Sequence in the Makespan, in Permutation Flow Shop Scheduling Heuristic Algorithms–An Analysis. *Indian Journal of Science and Technology*. 2016 Nov 17;9(42).

# Preserving Location Privacy in the IoT against Advanced Attacks using Deep Learning

Abdullah S. Alyousef<sup>1\*</sup>, Karthik Srinivasan<sup>2</sup>, Mohamad Shady Alrahhal<sup>3</sup>, Majdah Alshammari<sup>4</sup>, Mousa Al-Akhras<sup>5</sup>  
College of Computing and Informatics, Saudi Electronic University, Riyadh, Saudi Arabia<sup>1,2,5</sup>  
Department of Computer Science, King Abdulaziz University, Jeddah City, Saudi Arabia<sup>3</sup>  
Department of Computer Science, Hail University, Hail City, Saudi Arabia<sup>4</sup>  
Computer Information Systems Department, King Abdullah II School for Information Technology<sup>5</sup>  
The University of Jordan, Amman, Jordan<sup>5</sup>

**Abstract**—Location-based services (LBSs) have received a significant amount of recent attention from the research community due to their valuable benefits in various aspects of society. In addition, the dependency on LBS in the performance of daily tasks has increased dramatically, especially after the spread of the COVID-19 pandemic. LBS users use their real location to build LBS queries to take benefits. This makes location privacy vulnerable to attacks. The privacy issue is accentuated if the attacker is an LBS provider since all information about users is accessible. Moreover, the attacker can apply advanced attacks, such as map matching and semantic location attacks. In response to these issues, this work employs artificial intelligence to build a robust defense against advanced location privacy attacks. The key idea behind protecting the location privacy of LBS users is to generate smart dummy locations. Smart dummy locations are false locations with the same query probability as the real location, but they are far from both the real location and each other. Relying on the previous two conditions, the deep-learning-based intelligent finder ensures a high level of location privacy protection against advanced attacks. The attacker cannot recognize the dummies from the real location and cannot isolate the real location by a filtering process. In terms of entropy (the privacy protection metric), accuracy (the deep learning metric), and total execution time (the performance metric) and compared to the well-known DDA and BDA systems, the proposed system shows better results, where entropy = 15.9, accuracy = 9.9, and total execution time = 17 sec.

**Keywords**—LBS; dummy; deep-learning; attacks; accuracy; resistance; performance

## I. INTRODUCTION

The Internet of Things (IoT) can be defined as a network of devices that are connected through the Internet to facilitate performing tasks remotely. The IoT is involved in all aspects of people's lives, and it can be used in a wide range of applications in industry, transportation, and medicine [1]. In smart cities, the IoT forms the backbone for performing several missions, as shown in Fig. 1.

Among the IoTs, location-based services (LBSs) are considered the most important services that serve people daily. LBSs can be seen as commercial location applications that utilize the geographical location information of smart devices and mainly smartphones, enabling users to search for Points of Interest (PoIs), such as nearest restaurants, hospitals, libraries, and sports clubs [3]. In other words, LBSs employ a Global

Positioning System (GPS) to perform queries issued from the user side. In addition, smartphone users can easily obtain the benefits of LBS applications by downloading them from various sites, such as the Apple Store or Google Play Store. From an intersection of technologies point of view, LBS can be illustrated as shown in Fig. 2.

### A. The Importance of Location-based Services

In general, the importance of LBSs comes from their provided benefits, which make our lives easier and more enjoyable. In detail, three main sectors of daily life highlight the importance of LBS-enabled applications:

1) *Medical sector.* In the e-health field, LBSs play a significant role in monitoring patient health conditions (e.g. pulse rates and blood pressure levels), avoiding disasters [4, 5]. This, in turn, means that LBSs contribute to limiting the spread of illnesses such as COVID-19 by enabling medical staff and patients to avoid meeting and consequently maintaining a safe social distance.

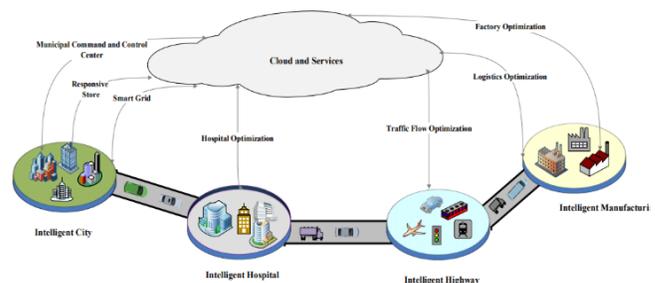


Fig. 1. IoT in Smart Cities [2].

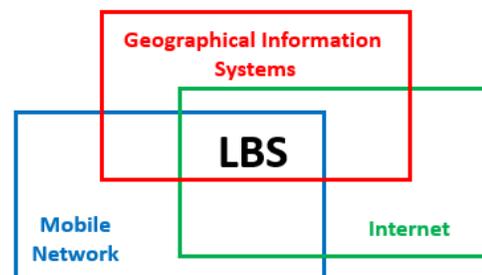


Fig. 2. LBS from an Intersection of Technologies Point of View.

\*Corresponding Author.

2) *Entertainment sector.* A further advantage of LBSs is enabling users to search for PIOs, such as nearby restaurants and music clubs, or enjoying games online [6, 7].

3) *Social sector.* Integrating LBS applications with wireless communication technologies have enabled the creation of location-based social networking services, such as Foursquare, Twinkle, and GeoLife [8]. This integration bridges the gap between the physical world and digital online social networking services.

**B. Statement of Problem**

The valuable benefits of the LBS applications mentioned above are not without risk. The key problem behind the extensive utilization of LBS applications is that the privacy of LBS users may be attacked. In the cybersecurity research field, privacy is a term that refers to sensitive information about users' interests, habits, or personal lives [9, 10]. Obtaining such information harms users and can even threaten their lives in cases of blackmailing or stealing valuable personal information, including the nature of their business, the details of their business trips, or their religious affiliation.

To gain a deep look at the privacy issue in LBS applications, the mechanism used for serving users should be analyzed. Using LBS applications requires constructing and sending queries relying on the real geographical locations of LBS users, who obtain their real locations through GPS. After manipulating these queries by the LBS provider, the results are returned to the users. Fig. 3 illustrates the general mechanism followed by LBS applications.

As shown in Fig. 3, there are three main steps, as follows.

- 1) The LBS user establishes a query using their real location. This query is then sent to the LBS provider.
- 2) The LBS provider processes the received query to answer the user. The result of the query (the retrieved POI) is packaged for resending.
- 3) The result is sent back to the LBS user and seen on the smartphone screen.

The scenario described in Fig. 3 is insecure against an attacker targeting the privacy of the LBS user. To define the problem accurately, modelling is required. Let  $\langle \alpha, \beta \rangle$  denote the coordinates of the real location of a given LBS user. Based on this representation, the query that is sent to the LBS provider is defined as:

$$Q_{LBS} = \{ \langle \alpha, \beta \rangle, S_{POI}, D, ID \}$$

Where:  $S_{POI}$ : set of points of interest that represent the result of the sent query.  $D$ : diameter of the search region (measured by Kilometres).  $ID$ : identity of the LBS user.

The privacy problem starts when an attacker tracks the real location of the LBS user or analyzes the sent query, as shown in Fig. 4.

In both cases (i.e., tracking the real location or analyzing the sent query), personal information about the LBS user is obtained. Malicious activities can be performed by a man-in-the-middle (MITM) attack. However, the privacy problem is

accentuated if the attacker is the LBS provider since all information is accessible. Upon this, the attacker (the LBS provider) can track the real location of the LBS user or analyze the received query. A malicious profile is then constructed on the attacker side, containing personal information that will be employed to attack the victim physically. Fig. 5 illustrates this dangerous scenario.

Formally, let  $VP$  denote the victim profile. Then,

$$VP = Track(\langle \alpha, \beta \rangle) \cup (Analyze Q_{LBS})$$

In terms of data flow and trust boundaries (attack surface), the security gap is represented by obtaining the query illustrated in Fig. 6.

It is worth mentioning that tracking the real location of LBS users leads to location privacy issues, and analyzing the sent query leads to query privacy issues [11]. In this work, we are concerned about location privacy only.

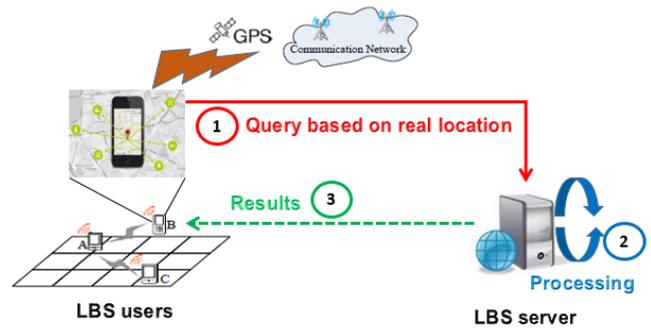


Fig. 3. The General Mechanism followed by LBS Applications.

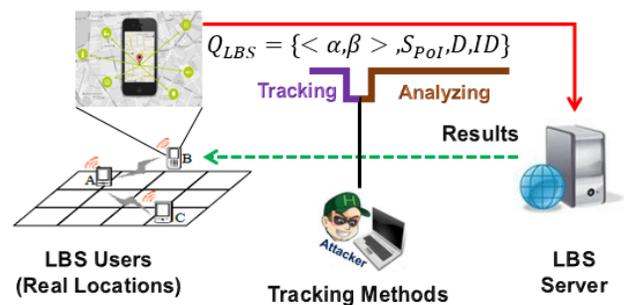


Fig. 4. Privacy Problem in LBS Applications.

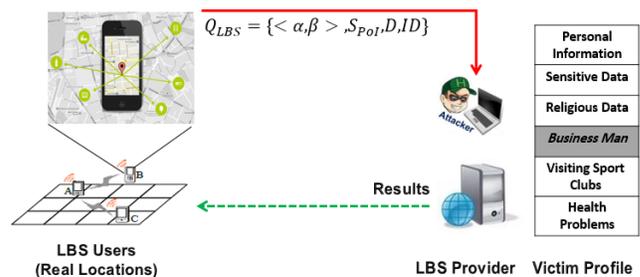


Fig. 5. Accentuated Privacy Problem in LBS Applications (LBS Provider is Attacker).

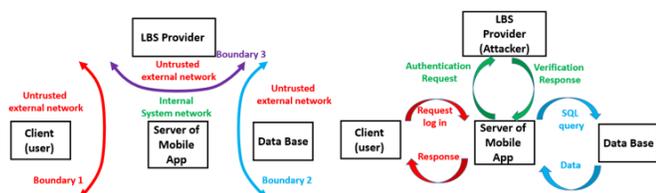


Fig. 6. Security Gap from the Attack Surface Perspective.

### C. Motivation and Research Questions

In light of the dangerous spread of the COVID-19 pandemic, dependence on mobile applications and the Internet has increased. This is because this dependency keeps people healthy in terms of achieving social distancing requirements. Increasing dependency on mobile applications is tightly coupled with an increasing level of privacy threats [12]. Moreover, existing advanced methods that could be used to track users, such as those that gather private information [13, 14], make privacy concerns more relevant. The capabilities of attackers are growing daily, with advanced attacks used to collect personal information from LBS users being applied. The attacker in a Map Matching Attack (MMA) employs the side information to gather sensitive data about the LBS user. In other words, the attacker can discover the kinds of activities the user is involved in by knowing the geographical map from which the LBS query is issued (i.e., without tracking the real location of the LBS user) [15, 16]. Fig. 7 illustrates the basic concept of an MMA.

Another advanced attack used for penetrating protection methods is the Semantic Location Attack (SLA) [18]. In an SLA, the attacker can infer semantic meanings related to the user's behavior, relying on both the time and place of where a user stays, as shown in Fig. 8.

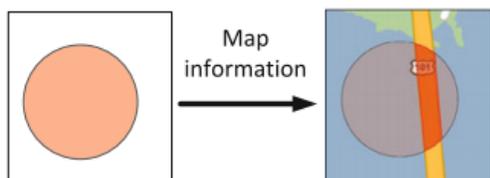


Fig. 7. Concept of MMA [17].

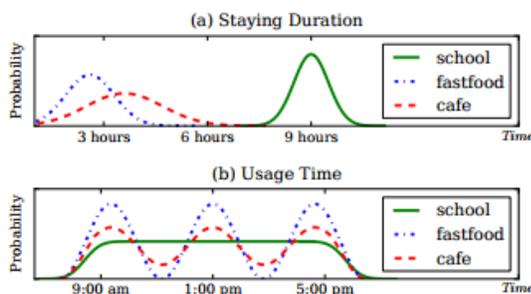


Fig. 8. Concept of SLA [19].

Motivated by these advanced attacks, two main research questions must be answered:

1) How do we ensure high resistance against both MMA and SLA?

2) How can privacy protection be quantified in terms of preventing attackers from penetrating privacy protection approaches?

### D. Contribution

The contributions of this work are listed as follows:

- In response to the first research question, a deep learning technique is proposed to generate strong dummy locations that protect the real location of the LBS user.
- In response to the second research question (second quality requirement), the entropy metric is employed to measure the resistance of the proposed deep learning based privacy protection system.

### E. Structure of the Work

The rest of this work is organized as follows. Related work is reviewed in Section II. Section III presents the methodology of designing and constructing the proposed system in detail. Security analysis is discussed in Section IV, followed by the results in Section V. Finally, the conclusion and suggestions for future work are provided in Section VI.

## II. RELATED WORK

In response to privacy concerns, researchers have proposed several approaches. The approaches were addressed from different perspectives, namely, server-based approaches, user-based approaches, and Trusted Third Party (TTP) approaches. Fig. 9 is a classification of LBS privacy protection approaches, where each category has its drawbacks.

The authors of work [20] proposed a Dummy Data Array (DDA) algorithm for generating dummy locations to protect the location privacy of LBS users. For a given region, which is divided into a grid of cells, the key idea of the DDA algorithm is to calculate both the vertices and the edges of each cell in the grid. Then, the DDA algorithm randomly selects some of the cells as dummy locations. To select strong dummy locations and achieve k-anonymity, the DDA algorithm selects k cells of equal area. The authors of the work [22] provided a survey of privacy protection approaches and they focused on dummies. Similarly, [21] uses dummies to protect the location privacy of LBS users, but with a different dummy generation method. The authors proposed two algorithms. The first is called CirDummy, which generates dummies based on a virtual circle that contains the real location of the LBS user. The second is called GridDummy, which generates dummies based on a virtual grid that covers the real location of the LBS user.

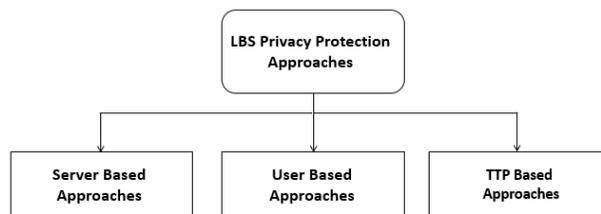


Fig. 9. Classification of Privacy Protection Approaches for LBS Applications.

Mix zones are defined in which all users' locations are hidden within these zones with some conditions to strengthen the protection method. In the work [23], the authors present the (DSC-LPP) approach to protect location privacy in the wireless channels. It is based on the idea of spatial cloaking-location privacy preserving. In the wireless channels, access points are essential elements in the structure of the network. Distance between the user and AP is the primary key for transforming and retrieving the location of users. Relying on this fact, the authors proposed to include the distance information in exchanged messages. The distance information can be exploited to confuse the attacker by manipulating it through mathematical transforming. The transformation leads to calculate a new location information, which in turn forms a clocking region. The clocking region reflects a security area of the location privacy. To enhance the performance, normalization is employed in the process of transformation for the purpose of building clocking regions. The advantage of this approach is that it can protect location privacy in a fixable way depending on increasing or decreasing the area of the clocking regions. However, if attackers have side information about the geographical map where the LBS user is located, the mechanism of protection becomes weak. In other words, this approach is not robust against MMA.

Pseudonym method [24] is used for protection of the user identity. The key idea is confusing the relationship between the position information and user identity information. This method is based on TTP model, TTP is the simplest intermediary entity between the user and the LBS provider. If the request is accepted, the request will be sent to the LBS provider; at the same time, the real ID will be changed to a pseudo-ID.

Information Retrieval (PIR) [25] was used to achieve full privacy protection. The key idea of the PIR technique depends on mathematical principle. It says that if it is impossible to compute a certain number or perform a certain mathematical task, then the information that form the task is protected. When the query is represented by a task, and PIR technique is applied, then, the LBS server can process and answer the query without knowing any sensitive information about the query.

### III. PROPOSED SYSTEM

This section is structured so that the threat model is defined first. The proposed system design is then described in detail. Next, security analyzes are discussed. Finally, the mechanism of evaluation of the proposed system is presented with the corresponding metrics.

#### A. Threat Model

The objective of the threat model is to draw the environment within which the proposed system is running and is expected to be robust against attackers. The threat model consists of four blocks as shown in Fig. 10.

- Attacker. The attacker is the LBS provider itself (or its maintainer), where all LBS queries are sent to it, and connecting with this malicious party is mandatory.
- Malicious goal. The goal of the attacker is to build a malicious profile about the LBS user. This is done by

gathering personal data about the victim through tracking the real locations used to establish LBS queries.

- Capabilities. The attacker's ability is supported by launching attacks on the victim, including MMA and SLA attacks.
- Type of attack. The type of each attack launched on the victim is active. This is because the LBS provider (attacker) can access all information received while serving the LBS user.

#### B. System Design

This section provides the architecture of the proposed system with its main components and the role of each component.

1) *Architecture of the proposed system:* The system decomposes three main components: the intelligent finder, query builder, and sender. The system is decentralized one because it is installed on each mobile device of LBS user. Table I summarizes the three components in terms of the assigned task, technique used, and installation.

Graphically, Fig. 11 shows the architecture of the proposed system with interconnections among the three components.

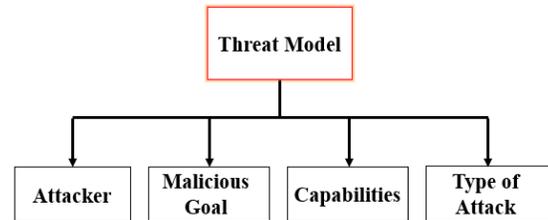


Fig. 10. Blocks of the Threat Model.

TABLE I. COMPONENTS OF THE SYSTEM

| Name               | Task                         | Technique                                                        | Installation          |
|--------------------|------------------------------|------------------------------------------------------------------|-----------------------|
| Intelligent finder | Generating dummy locations   | Convolutional Neural Network (CNN), Support Vector Machine (SVM) | LBS user (Smartphone) |
| Query builder      | Building the protected query | Anonymity of identity                                            | LBS user (Smartphone) |
| Sender             | Sending the protected query  | Wireless communication                                           | LBS user (Smartphone) |

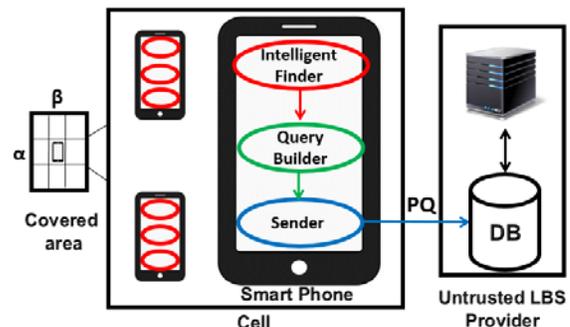


Fig. 11. Architecture of the Proposed System.

As shown in Fig. 11, the smartphone (which represents an LBS user) is located in a certain location (cell) within the covered area that consists of  $(\alpha \times \beta)$  cells. The rest of the cells are spread on different regions that contain various PoIs. The cells that form the covered area can be exploited as dummy locations to protect the location privacy of LBS users. In other words, using fake locations instead of the real location cuts the tracking series that is performed on the attacker's side to complete the malicious profile. This is because the attacker (LBS provider) cannot recognize the real location among dummies. However, the attacker attempts to compromise the protection method by applying advanced attacks such as MMA and SLA. This requires that the process of generating (or finding) dummy locations be accurate to provide strong dummies that can protect location privacy against advanced attacks. In this work, artificial intelligence is employed to generate strong dummies based on Convolutional Neural Networks (CNNs) and Support Vector Machines (SVMs). Upon this, the intelligent finder selects (generates, or searches) strong dummy locations and then provides them to the query builder to establish a query with multiple locations (one of them is the real one). Then, the sender sends the protected query (PQ) to the LBS provider (attacker). The attacker is confused about determining the real location among dummies. Below is a detailed description of the role of each component.

2) *Role of an intelligent finder*: The main task of this component is ensuring the location privacy of the LBS user. This is performed by protecting the location information used to form the sent query. Based on a novel location privacy protection approach, namely, Vectors of Protection (VoP), this component ends its assigned task. VoP fills a vector of locations by dummies, and the real location in the LBS query is replaced by this vector. The key idea of the VoP approach is illustrated in Fig. 12.

As shown in Fig. 12, the real location of the LBS user ( $LBS_L^r$ ) is represented by the left side. The role of the intelligent finder component is to fill the vector by dummy locations by executing the VoP approach. The rest of the query units remain constant.

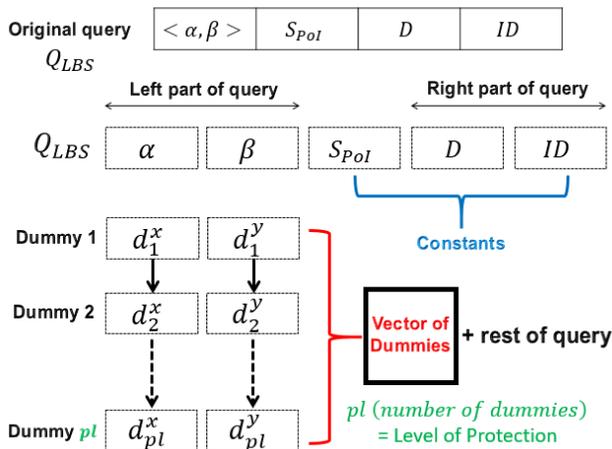


Fig. 12. Key idea of the VoP Approach.

In detail, for a region divided into  $\alpha \times \beta$  cells, the real location of the LBS user  $LBS_L^r$  is located in a certain cell. Each cell has a query probability  $CELL_p^q$ . The query probability is a term that refers to the number of queries sent from a specific location in the past (i.e., number of queries built based on the cell divided by the total number of queries built based on the whole cells). Each cell has a certain value of query probability, as shown in Fig. 13.

The VoP approach selects dummies randomly. From the real location of the LBS user, some vectors are issued to the selected dummies. Then, the selected dummies are stored in the vector of dummies. The number of dummies determines the level of protection. This means that the LBS user has full control over the desired level of privacy protection. For instance, if the LBS user selects 3 dummy locations, the level of privacy protection is 4. This is because the real location is surrounded by three dummies, as shown in Fig. 14.

The process of selecting dummies randomly without any constraint is a poor tactic. This is because the query probability of each dummy location differs from the query probability of the real location of the LBS user. This increases the ability of the attacker to determine the real location among dummies. Therefore, it is better to select dummy locations with the same query probabilities as the real location of the LBS user. Fig. 15 illustrates the selection process under the same query probability condition.

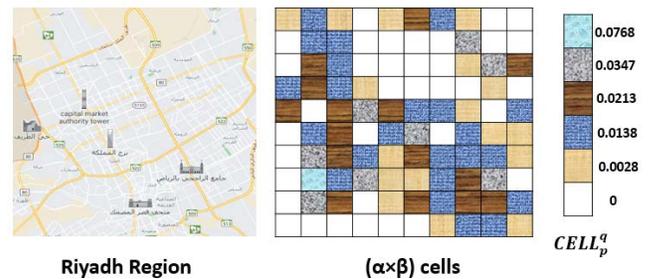


Fig. 13. Query Probabilities of Cells.

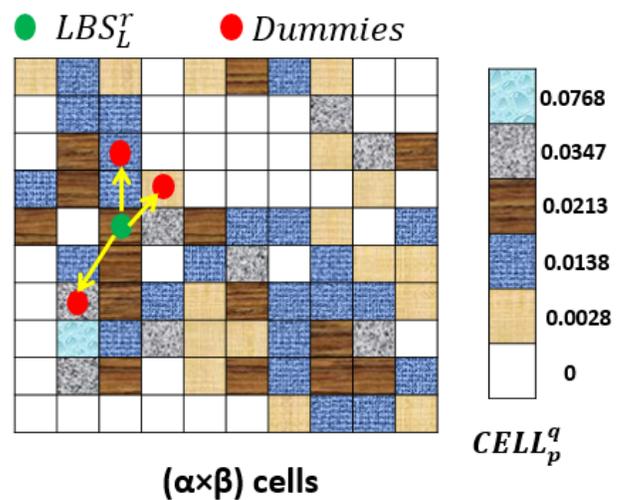


Fig. 14. Achieving Privacy Protection of 4 Levels.

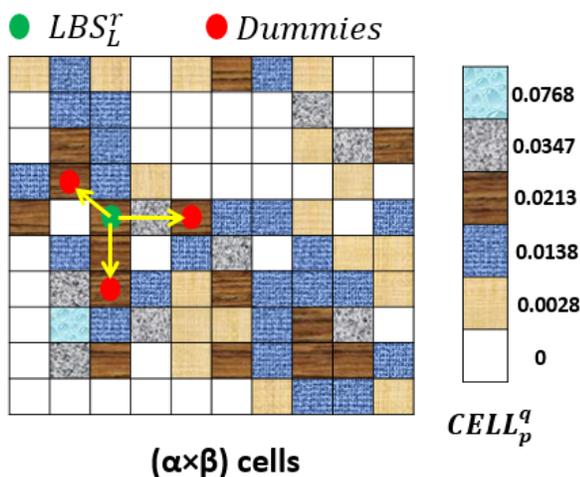


Fig. 15. Selecting Dummies Depending on the Same Query Probability Condition.

This method guarantees that the uncertainty (at the attacker's side) in determining the real location among the dummies is maximal. Mathematically, this uncertainty is represented by entropy. Entropy is a term that refers to the inability to determine an object among others based on the same features [26]. The entropy of identifying the real location out of the dummy vector  $ENT_{dv}^r$  is defined as:

$$ENT_{dv}^r = - \sum_{i=1}^{pl} CELL_{p_i}^q \times \log_2 \times CELL_{p_i}^q \quad (3)$$

where  $pl$  denotes the protection level of privacy.

Despite selecting dummies based on the query probabilities condition, the privacy threat remains. Selecting weak dummy locations creates a vulnerability where the attacker can apply MMA and SLA successfully. Weak dummies mean that the dummy locations are near the real location of the LBS user. This allows the attacker to filter dummies easily. This requires additional conditions in the process of selecting dummy locations. This condition states that the selected dummies must be far away from the real location of the LBS user, as shown in Fig. 16.

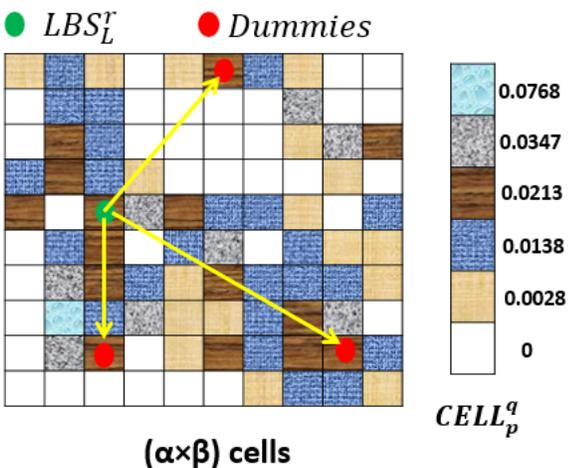


Fig. 16. Selecting dummies Depending on both the same Query Probabilities and Far away Conditions.

The actual selection process is performed by the intelligent finder component. Electing suitable dummies (i.e., strong dummies) requires an intelligent method. This intelligent method depends on scanning the covered region and then determining strong dummies. In this work, a deep learning method (the CNN network) with the help of SVM is employed to elect strong dummies to fill the vector of dummies.

The task of the CNN is extracting the features of a given geographic region (map). This is performed by scanning the map through two kinds of layers: convolutional layers and pooling layers. Fig. 17 illustrates the mechanism used by the CNN for extracting features of a given region.

As shown in Fig. 17, the CNN goes through the map in a convolutional manner (depending on a filter or kernel) to extract the first level of features. Then, the extracted features are grouped through a pooling layer to draw a deep look at the locations included in the map. This procedure (i.e., convolution and pooling) is repeated frequently for the series of extractions. The final pooling layer includes the final features. Among the extracted features, some locations are suitable to be strong dummies, while some are weak dummies. The SVM is linked to the fully connected layer to classify the locations into two main groups: strong dummies and weak dummies. SVM is an intelligent technique that separates a given set of data into two major classes. SVM relies on margin, which can be seen as a restricted area between the two classes. Fig. 18 shows the basic concept of SVM.

SVM is represented mathematically by the sigmoid function, which forms the (S) curve from a graphical perspective, as shown in Fig. 19.

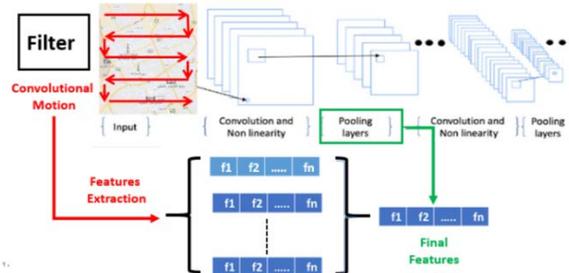


Fig. 17. Extracting Features of Map using CNN.

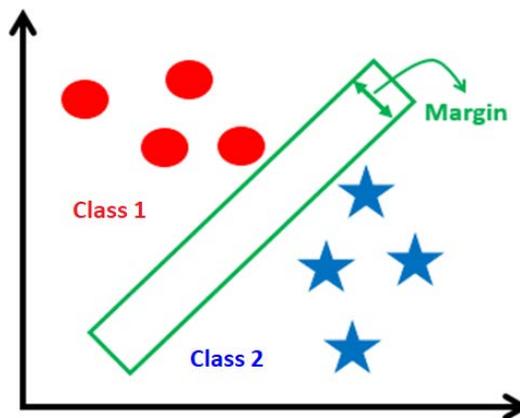


Fig. 18. Basic Concept of SVM.

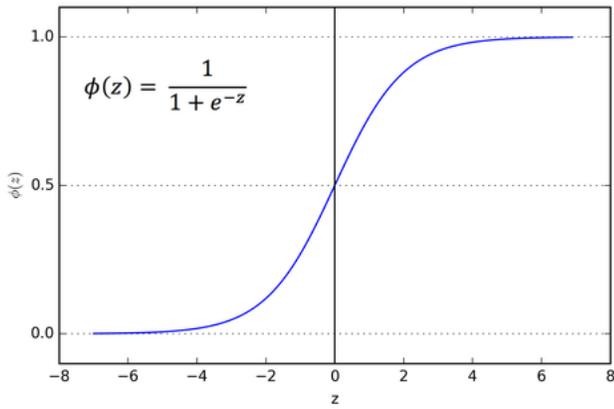


Fig. 19. Sigmoid Function.

According to the sigmoid function, the complete CNN network will be seen as a classifier, where all values above (+) or lower than (0) represent strong dummies, and the area between the range [0, +1] represents the margin. Fig. 20 shows the complete CNN.

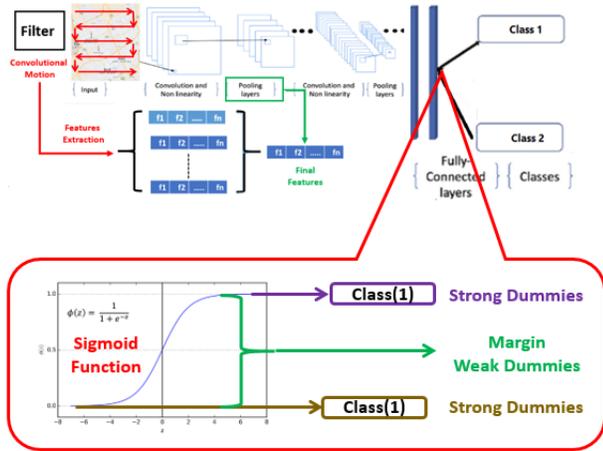


Fig. 20. Complete CNN Network.

Regarding the VoP approach, the complete CNN feeds it with a set of strong dummies. Let  $final_{CNN}^{dummies}$  denotes the final set of strong dummies. The vector of dummies is then filled by a random selection of dummies since they all satisfy the two conditions. The size of the VoP (or the number of selected dummies) is based on the privacy protection level that is desired. Mathematically,

$$VoP = random \{ final_{CNN}^{dummies} \} \quad (4)$$

The intelligent finder represented by the CNN is trained on the Brightkite dataset [27]. It consists of 7.3 million rows and five columns (user ID, check-in time, latitude, longitude, and location id). To involve query probabilities, we add a new column QP to the database. The values of the QP are generated randomly. Additionally, the Brightkite dataset is used for the testing stage. Therefore, the dataset is divided into two parts, as shown in Fig. 21.

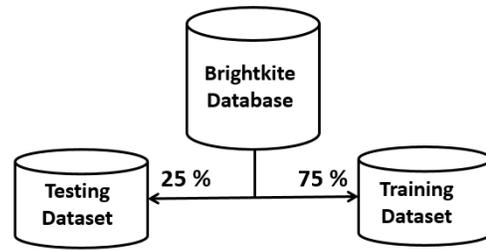


Fig. 21. Dividing the Brightkite Dataset.

3) *Role of the query builder:* This component is responsible for building the protected query. It receives the vector of dummies generated by executing the VoP approach (the task of the intelligent finder) and then constructs the query. To add a second layer of privacy protection, another task is assigned to this component, which blurs the ID of the LBS user. To end this, the query builder components use an anonymity technique. The key idea behind the anonymity technique is to hide the ID of the LBS user by replacing it with a fake ID. Upon this, the query generated by the query builder component is constructed as shown in Fig. 22.

As shown in Fig. 22, the ID of the LBS user in the original query is replaced by a fake identity ( $\overline{ID}$ ). This adds additional protection to location privacy since the attacker (LBS provider) can recognize neither the real location among dummies nor the identity of the LBS user. Thus, the whole units of the protected query are given by:

$$Protected[Q_{LBS}] = \{ \langle \begin{matrix} \alpha & \beta \\ d_1^x & d_1^y \\ d_2^x & d_2^y \\ \vdots & \vdots \\ d_{pl}^x & d_{pl}^y \end{matrix} \rangle, S_{Pol}, D, \overline{ID} \} \quad (5)$$

4) *Role of the sender:* This component is responsible for sending the protected LBS query to the LBS provider for manipulation. Fig. 23 illustrates the task of the sender component.

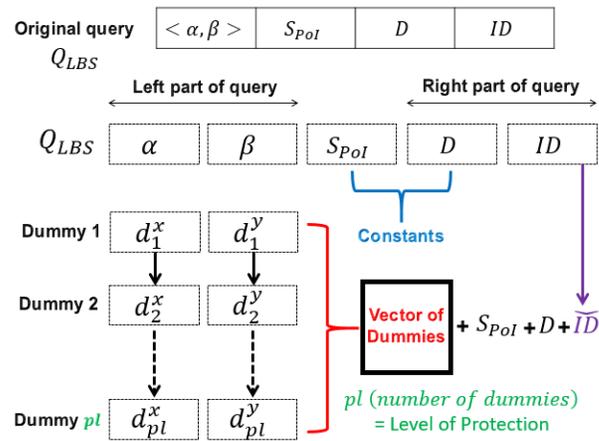


Fig. 22. The Query Builder Constructing a Protected Query.

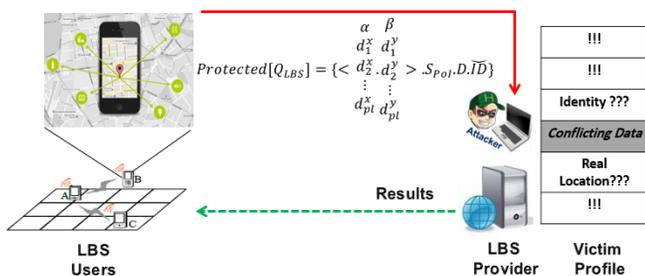


Fig. 23. Sending a Protected Query to the LBS Provider (Attacker).

Manipulating the protected LBS query at the LBS provider-side leads to confusion. This is because the victim profile will be full of conflicting data that is not beneficial. Thus, the location privacy of the LBS user is protected against the LBS provider if it acts maliciously.

#### IV. SECURITY ANALYSIS

This section discusses the security analyses, where the issues are taken from the attacker's perspective. In this context, two main issues are involved, as shown in Fig. 24.

In regard to knowing the proposed approach by the attacker, a reversing trial is expected to be performed to filter dummies. Here, the power of the randomization process is employed. As mentioned in formula 4 above, the final set of dummies is selected randomly. This means that if 20 dummies are strong and can be used as actual dummies, 5 dummies can be selected randomly to be utilized as actual dummies to achieve a protection level of 6 degrees. Randomization ensures complete doubt about determining which of the 5 dummies are elected among the 20 dummies. This reflects uncertainty in the process of selecting dummies on the attacker side. As a result, the attacker can only randomly guess the real location among dummies. Thus, reversing the VoP approach fails to achieve the malicious goal of the attacker.

In regard to discussing the success of the MMA and SLA attacks from the perspective of the attacker, these attacks fail. The reason is that the VoP approach takes into account the  $CELL_p^q$ , where it is the same for all locations (the real and dummies). In addition, the dummies selected by the CNN are far from both each other and the real location. Therefore, attempting to collect the dummies in one region is inapplicable at the attacker side. This means that the success of the attacks will not be achieved.

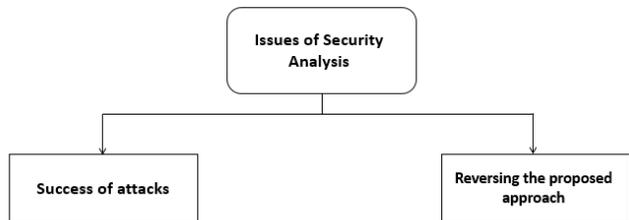


Fig. 24. Issues of Security Analysis.

#### V. RESULT AND DISCUSSION

This section provides the results in the context of comparison with two approaches. The first approach is the classical one. A classical approach is an approach inspired by the proposed VoP approach, where the dummies are selected randomly without taking any optimization into account (i.e.,  $CELL_p^q$  and distance-relation between the  $LBS_L^r$  and the other dummies). The classical approach is referred to as the basic dummy approach (BDA). The second approach involved in the comparison is the one that is proposed in ref [20], which is DDA.

##### A. Evaluation Metrics

In this work, three types of metrics are used for evaluation, as shown in Fig. 25.

The privacy-based metric, entropy, which is defined above (by formula 3), is employed to measure the privacy protection level that is achieved. Entropy is a metric addressed by many authors who conducted surveys, such as [28- 32], and by others who made technical research papers, such as [33-36]. The mechanism of an evaluation relying on the entropy metric is adjusted by the following rules:

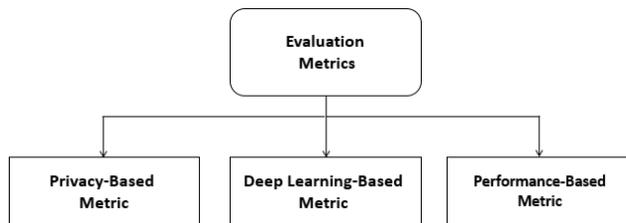


Fig. 25. Types of Evaluation Metrics.

- 1) There is no upper limit to the value of entropy.
- 2) There is no lower limit to the value of entropy.
- 3) A higher entropy value means a higher privacy protection degree.
- 4) A lower entropy value means a lower privacy protection degree.

For the deep-learning-based metric, an accuracy metric is utilized. Accuracy is a term that refers to the ratio of records (dummies) that are correctly classified (i.e., selected as strong dummies) [37]. The accuracy metric is inspired by a confusion matrix, a common term in the data mining research field [38]. Table II shows the confusion matrix (COFMX).

TABLE II. COFMX AND ITS COMPONENTS

| Actual dummy (Predicted dummy) | Confusion matrix          |                           |                 |
|--------------------------------|---------------------------|---------------------------|-----------------|
|                                | DM                        | $\neg$ DM                 | Sum             |
| DM                             | True positives DM (TPDM)  | False negatives DM (FNDM) | TPDM + FNDM = P |
| $\neg$ DM                      | False positives DM (FPDM) | True negatives DM (TNDM)  | FPDM + TNDM = N |

Where:

- 1) TPDM is a positive dummy that is correctly labelled by the CNN classifier.
- 2) TNDM is a negative dummy that is correctly labelled by the CNN classifier.
- 3) FPDM is a negative dummy that is incorrectly labelled positive.
- 4) FNDM is a positive dummy that is mislabelled negative.

Accuracy is given by the following formula:

$$Accuracy = \frac{(TPDM+TNDM)}{\text{number of all records/dummies in the testing set}} \quad (6)$$

The mechanism of evaluation relying on the accuracy metric is adjusted by the following rules:

- 1) There is an upper limit to the value of accuracy (1 or 100%).
- 2) There is a lower limit to the value of accuracy (0).
- 3) A higher accuracy value means a higher prediction degree.
- 4) Lower accuracy value means a lower prediction degree.

For the performance-based metric, time dominates the case. Thus, the total execution time ( $TexeT$ ) required to execute the approach is used. The  $TexeT$  is defined by:

$$TexeT = T_{VoP}^{exe} + 2 \times T_{query}^{send} + T_{query}^{processing} \quad (7)$$

where  $T_{VoP}^{exe}$  refers to the time of executing the proposed VoP approach at the smartphone of the user,  $T_{query}^{send}$  refers to the sending time of the query (assuming that the return takes the same time), and  $T_{query}^{processing}$  refers to the processing time at the server-side to answer the send query. The mechanism of evaluation relying on the  $TexeT$  metric is adjusted by the following rules:

- 1) There is no upper limit to the value of  $TexeT$ .
- 2) There is no lower limit to the value of  $TexeT$ .
- 3) A higher  $TexeT$  value means a lower performance degree.
- 4) A lower  $TexeT$  value means a higher performance degree.

### B. Entropy-Based Evaluation without Threats

Without applying any threat, the value of entropy is calculated to increase the protection level from 3 to 21 (i.e., from three dummies to 21 dummies). Fig. 26 shows the results.

Discussion and justifications: As shown in Fig. 26, the CNN-VoP and BDA approach experiences increased entropy values as the protection level increases. The reason is related to mathematical justification, where increasing the number of dummies involved in the protection level leads to an increase in the entropy value. However, the CNN-VoP approach achieves better scores than the BDA approach. This is due to selecting dummy locations under the control of  $CELL_p^q$  in the CNN-VoP approach. In contrast, no constraints are used in the BDA approach. For the DDA approach, its curve can be divided into three parts. The first part behaves the same as the

CNN-VoP and BDA approaches to increase the values of entropy. In the first part, the DDA sometimes overcomes the BDA depending on the tree that combines similar dummies, which in turn means that some dummies have  $CELL_p^q$  that are similar to the  $CELL_p^q$  of the real location or approximately close to it. In the second part, where PL=12, the DDA performs the worst. This is because there are no available candidates that can be used as actual dummies, and in this case, the DDA repeats the dummies, which negatively affects the entropy value. In the third part, the DDA enhances slightly, but the BDA outperforms it due to the broad set of dummies available compared to a limited set controlled by the DDA.

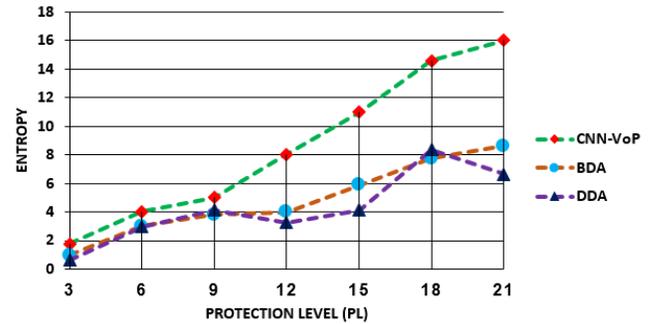


Fig. 26. Value of Entropy Metric vs. Increasing PL without any Threat.

### C. Entropy-Based Evaluation under Threats

After applying the MMA threat, the value of entropy experiences a decreasing trend compared to the normal situation; the attacker has no information about the geographic map from which LBS queries are sent. This is shown in Fig. 27.

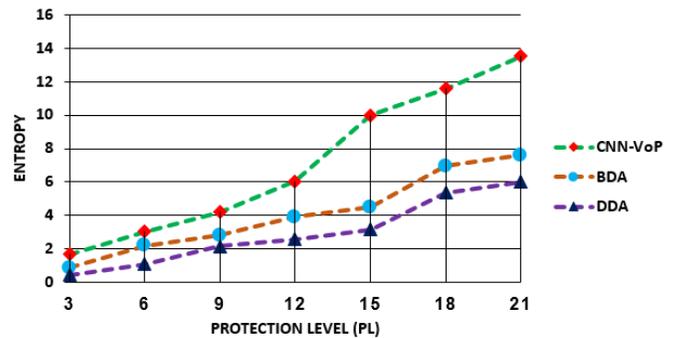


Fig. 27. Value of Entropy Metric vs. Increasing PL under MMA Threat.

Discussion and justifications: As shown in Fig. 27, despite the negative impact of the MMA threat, the CNN-VoP approach maintained its peak position. This is because of the factors taken into account in the procedure for selecting the dummies, where (1) the  $CELL_p^q$  of each dummy is the same as the real location (which contributes to destroying the benefits that may be gained at the attacker side by analyzing the dummies if they are located in well-known areas) and (2) the dummies are spread over a wide space that cannot be collected in one area for malicious filtering by the attacker. The BDA scheme overcomes the DDA scheme since DDA is vulnerable to selecting the dummies based on area similarities. This gap can be exploited by the attacker to filter some dummies,

weakening the defense that is created by the DDA. In contrast, the BDA scheme ignores the similarities since it selects dummies randomly, avoiding the drawback of the DDA scheme. Table III shows the numerical results of the entropy metric.

The results summarized in Table III show that the BDA scheme experiences slight weaknesses against MMA attacks compared to a significant weakness in the DDA scheme.

For robustness against the SLA attack, Fig. 28 shows the documented results, where a severe negative impact is clearly seen in the DDA approach compared to the normal situation.

Discussion and justifications: As shown in Fig. 28, the common fact that "the value of entropy increases as the PL increases" is still working in all schemes involved in comparison. However, the entropy values are less when compared to the values under MMA threat, which reflects that the SLA is more dangerous than the MMA threat. Despite this change, the CNN-VoP scheme is still ranked at the top, followed by the BDA scheme. At the last position, the DDA scheme is coming. This scenario can be justified by the positive contribution of using CNN to scan and elect strong dummies for membership in the final set of dummies used for privacy protection. The BDA scheme ignores the factor related to elect dummies that achieve the condition of distance (i.e., long distances between the elected dummies and the real location). The DDA scheme employs none of the factors, and therefore, it performs the worst as a protection method. Table IV shows the entropy values after applying the SLA threat.

To address the difference between the MMA and SLA threats, Fig. 29 shows a visual representation of the entropy values documented in Table III and Table IV.

Table V shows the transformation of the visual representation of Fig. 29 into numeric values.

TABLE III. RESULTS OF ENTROPY IN THE THREE SCHEMES UNDER MMA THREAT

| Approach \ PL |                | 3     | 6     | 9     | 12    | 15    | 18     | 21     |
|---------------|----------------|-------|-------|-------|-------|-------|--------|--------|
| CNN-VoP       | Entropy Values | 1.658 | 3.046 | 4.208 | 6.046 | 9.987 | 11.587 | 13.508 |
| BDA           |                | 0.854 | 2.196 | 2.809 | 3.906 | 4.501 | 6.946  | 7.609  |
| DDA           |                | 0.427 | 1.078 | 2.145 | 2.578 | 3.150 | 5.347  | 6.005  |

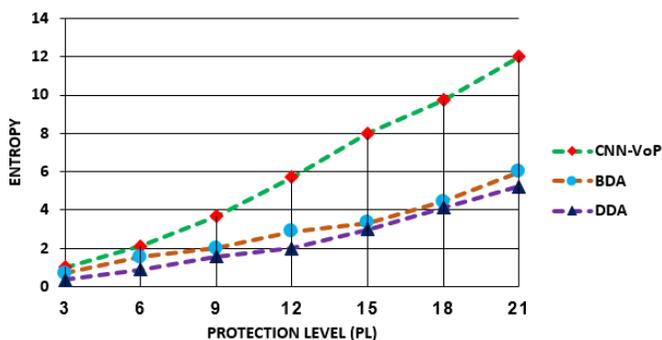


Fig. 28. Value of Entropy Metric vs. Increasing PL under SLA Threat.

TABLE IV. RESULTS OF ENTROPY IN THE THREE SCHEMES UNDER SLA THREAT

| Approach \ PL |                | 3     | 6     | 9     | 12    | 15    | 18    | 21     |
|---------------|----------------|-------|-------|-------|-------|-------|-------|--------|
| CNN-VoP       | Entropy Values | 1.007 | 2.113 | 3.666 | 5.711 | 7.999 | 9.720 | 11.994 |
| BDA           |                | 0.700 | 1.539 | 1.999 | 2.878 | 3.332 | 4.448 | 5.996  |
| DDA           |                | 0.364 | 0.886 | 1.589 | 1.988 | 2.997 | 4.123 | 5.231  |

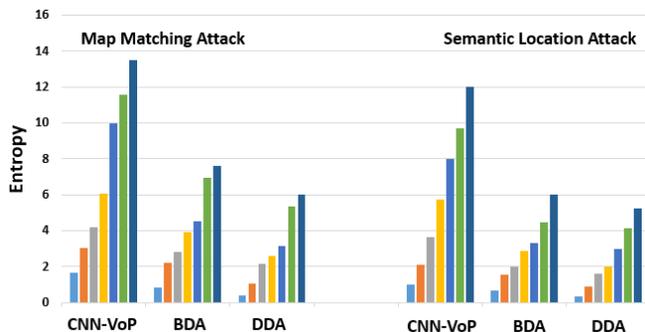


Fig. 29. Visual Representation of the Entropy values under MMA and SLA Threats.

TABLE V. DIFFERENCE IN ENTROPY VALUES AMONG THE THREE SCHEMES AFTER APPLYING MMA AND SLA THREATS

| Approach \ PL |                              | 3     | 6     | 9     | 12    | 15    | 18    | 21    |
|---------------|------------------------------|-------|-------|-------|-------|-------|-------|-------|
| CNN-VoP       | Difference of Entropy Values | 0.651 | 0.933 | 0.542 | 0.335 | 1.988 | 1.867 | 1.514 |
| BDA           |                              | 0.154 | 0.657 | 0.81  | 1.028 | 1.169 | 2.498 | 1.613 |
| DDA           |                              | 0.063 | 0.192 | 0.556 | 0.579 | 0.153 | 1.224 | 0.774 |

Table V shows that the SLA threat has a more negative impact on the privacy of LBS users than the MMA threat. This is because the attacker employs time usage and knowledge about the geographic map to attack privacy (or filter some dummies). Thus, it is recommended to pay more attention to the semantic location threat in the location privacy research arena.

#### D. Accuracy-Based Evaluation

For the accuracy of electing suitable (or strong dummy locations), Fig. 30 illustrates the output of the three schemes.

Discussion and justifications: As shown in Fig. 30, the CNN-VoP scheme performs the best, followed by the BDA and DDA schemes. The root reason for this is related to using SVM as a classifier in the structure of the CNN used to scan and discover the dummy locations. Due to the series of convolutional and pooling layers used in the CNN, effective features of the region that includes the real location of the LBS user are generated. Based on the extracted features, strong dummies that satisfy the two conditions are elected. This means that some strong base criteria are used in the CNN-VoP scheme compared to poor ones in the other two schemes. This helps to add another strong justification about the strength of the CNN-VoP scheme in deep learning. This, in turn, provides proof of why entropy values are higher in both cases (i.e., without a threat and under the threat of attack) discussed above.

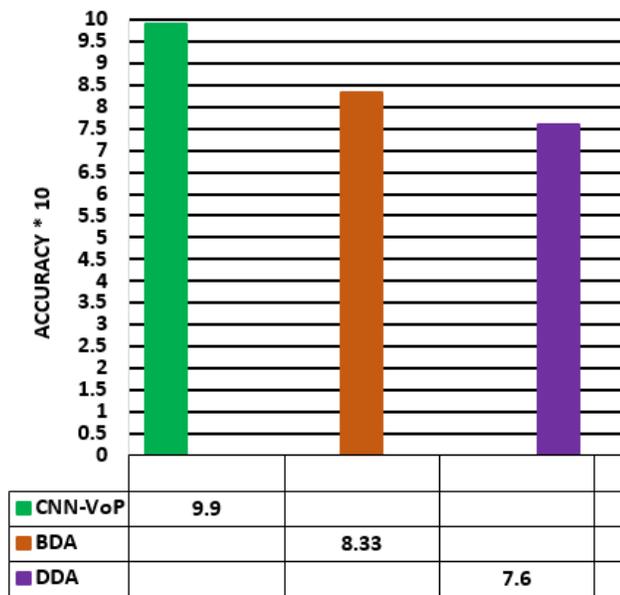


Fig. 30. Accuracy of Three Schemes

### E. Performance-Based Evaluation

According to the increased number of sent  $Q_{LBS}$ , the total execution time is calculated for the three schemes, as shown in Fig. 31.

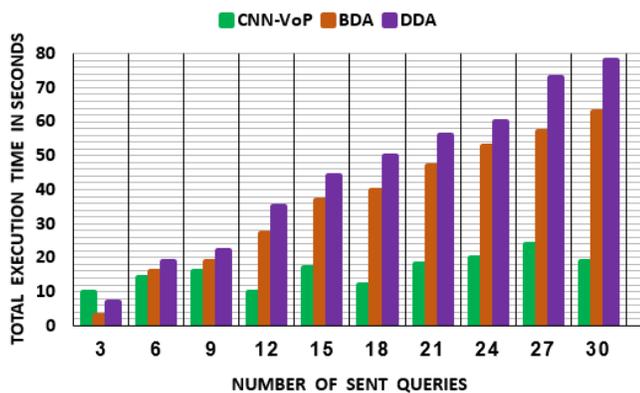


Fig. 31. Performance of Three Schemes in Terms of  $TexeT$ .

Discussion and justifications: As shown in Fig. 31, the CNN-VoP scheme performs the best compared to the BDA and DDA schemes, except at the beginning, where it is the worst. The reason that the CNN-VoP scheme performs the worst at the beginning is related to the training stage, where it still has no knowledge and requires much more time. During this period, the BDA scheme selects the dummies randomly, which is a fast method and leads to the shortest time. The DDA scheme consumes some time to construct the tree of dummies and thus comes after the BDA. After that, when increasing the number of sent queries,  $TexeT$  increases in all schemes. However, the CNN-VoP scheme achieves the best performance since, after the training completes, it can select dummies directly based on a knowledge database. The BDA scheme returns to the second-order, while the DDA performs the worst (the longest time). This is related to a common problem when using a decision tree, which is an overfitting problem. This

means that the process of constructing the decision tree requires manipulating all branches without ignoring any branches. This consumes substantial time and leads to poor performance by the DDA scheme. In terms of average,  $TexeT = 17 \text{ sec}$  for the CNN-VoP scheme,  $TexeT = 47 \text{ sec}$  for the BDA scheme, and  $TexeT = 67 \text{ sec}$  for the DDA scheme.

### VI. CONCLUSION AND FUTURE WORK

Recently, the world witnessed a widespread COVID-19 pandemic which changed the way people performed daily tasks. In this context, and to avoid infection, people tended to use location-based services (LBSs), which have received great attention from companies and research groups. Relying on an LBS opens the door for attackers to attack the privacy of LBS users since performing tasks requires sending the user's real location. The problem is accentuated concerning advanced methods that attackers can use, such as Map Matching Attacks (MMAs) and Semantic Location Attacks (SLAs). The privacy of LBS users will be under great threat if the LBS provider acts as an attacker and can apply MMA and SLA attacks. In responding to this challenge, this work presents a location privacy protection system. The system consists of three main components. The first component is the intelligent finder. The role of the intelligent finder is to find (or select) strong dummy locations for privacy protection against the malicious party (the LBS provider), such that the attacker will be confused about determining the real location of the LBS user among the dummies. The intelligent finder uses a deep learning technique, which is the Convolutional Neural Network (CNN). The CNN is employed to create a classifier that classifies locations found in the region where the LBS user is located into the categories of weak and strong dummies. After creating the strong dummy category, a Vector of Protection (VoP) approach is performed. Strong dummies satisfy two main constraints: (1) the query probability of each selected dummy is the same as the real location, and (2) they are spread away from each other and the real location. The previous two constraints ensure high resistance against advanced MMA and SLA threats. The second component is the query builder, which is responsible for (1) constructing the protected query based on the selected strong dummies and (2) hiding the identity of the LBS user. The third component is the sender, which is responsible for sending the protected query to the LBS provider. The proposed location privacy protection system is evaluated according to entropy (the privacy protection metric), accuracy (the deep learning metric), and total execution time (the performance metric). Compared to well-known systems, which are the DDA and the BDA, the proposed system shows better results, where entropy = 15.9, accuracy = 9.9, and total execution time = 17 sec.

Limitation: Privacy protection for an LBS considers location privacy and query privacy; the sent query can be analyzed depending on the query sampling attack. In attacking query privacy, the attacker relies on the PoI as well as its link with the locations. In this work, query privacy was not taken into consideration.

Future work: In future work, we will enhance the proposed system to ensure comprehensive privacy protection in LBS

applications (i.e., ensuring both location privacy and query privacy). In addition, we will test the system using different databases for training the intelligent finder and use another advanced intelligent method, such as advanced clustering.

#### REFERENCES

- [1] Aceto, Giuseppe, Valerio Persico, and Antonio Pescapé. "Industry 4.0 and health: Internet of things, big data, and cloud computing for healthcare 4.0." *Journal of Industrial Information Integration* 18 (2020): 100129.
- [2] Kassab, Wafa'A., and Khalid A. Darabkh. "A-Z survey of Internet of Things: Architectures, protocols, applications, recent advances, future directions and recommendations." *Journal of Network and Computer Applications* 163 (2020): 102663.
- [3] Jiang, Hongbo, et al. "Location Privacy-preserving Mechanisms in Location-based Services: A Comprehensive Survey." *ACM Computing Surveys (CSUR)* 54.1 (2021): 1-36.
- [4] Garg, Niharika. "Technology in Healthcare: Vision of Smart Hospitals." *Handbook of Research on Engineering, Business, and Healthcare Applications of Data Science and Analytics*. IGI Global, 2021. 346-362.
- [5] Ma, Yujun, et al. "Big health application system based on health internet of things and big data." *IEEE Access* 5 (2016): 7885-7897.
- [6] Girginkaya Akdağ, Suzan, and Ahu Ergen. "Role of location-based mobile apps in city marketing: Beşiktaş as a student-friendly district." *Journal of Location Based Services* 14.2 (2020): 49-70.
- [7] Uphaus, PerOle, et al. "Location-based services—the market: success factors and emerging trends from an exploratory approach." *Journal of Location Based Services* (2021): 1-26.
- [8] Wawrowski, Bartosz, and Iwona Otolá. "Social Media Marketing in Creative Industries: How to Use Social Media Marketing to Promote Computer Games?." *Information* 11.5 (2020): 242.
- [9] Zou, Shihong, et al. "CrowdHB: A Decentralized Location Privacy-Preserving Crowdsensing System Based on a Hybrid Blockchain Network." *IEEE Internet of Things Journal* (2021).
- [10] Jiang, Hongbo, et al. "Location Privacy-preserving Mechanisms in Location-based Services: A Comprehensive Survey." *ACM Computing Surveys (CSUR)* 54.1 (2021): 1-36.
- [11] Almusaylim, Zahrah A., and N. Z. Jhanjhi. "Comprehensive review: Privacy protection of user in location-aware services of mobile cloud computing." *Wireless Personal Communications* 111.1 (2020): 541-564.
- [12] Lv, Wenzhe, et al. "Towards Large-Scale and Privacy-Preserving Contact Tracing in COVID-19 pandemic: A Blockchain Perspective." *IEEE Transactions on Network Science and Engineering* (2020).
- [13] Dardari, Davide, Pau Closas, and Petar M. Djurić. "Indoor tracking: Theory, methods, and technologies." *IEEE Transactions on Vehicular Technology* 64.4 (2015): 1263-1278.
- [14] Zhang, Lan, et al. "Montage: Combine frames with movement continuity for realtime multi-user tracking." *IEEE Transactions on Mobile Computing* 16.4 (2017): 1019-1031.
- [15] Liu, Ting, et al. "User Personalized Location k Anonymity Privacy Protection Scheme with Controllable Service Quality." *International Conference on Machine Learning for Cyber Security*. Springer, Cham, 2020.
- [16] Jagwani, Priti, and Saroj Kaushik. "Privacy in location based services: Protection strategies, attack models and open challenges." *International conference on information science and applications*. Springer, Singapore, 2017.
- [17] Wernke, Marius, et al. "A classification of location privacy attacks and approaches." *Personal and ubiquitous computing* 18.1 (2014): 163-175.
- [18] Kuang, Li, et al. "Using location semantics to realize personalized road network location privacy protection." *EURASIP Journal on Wireless Communications and Networking* 2020.1 (2020): 1-16.
- [19] Lee, Byoungyoung, et al. "Protecting location privacy using location semantics." *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011.
- [20] Alrahhal, Mohamad Shady, et al. "AES-Route Server Model for Location based Services in Road Networks." *International Journal of Advanced Computer Science And Applications* 8.8 (2017): 361-368.
- [21] Lu, Hua, Christian S. Jensen, and Man Lung Yiu. "Pad: privacy-area aware, dummy-based location privacy in mobile services." In *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, pp. 16-23. ACM, 2008.
- [22] Wernke, M., Skvortsov, P., Dürr, F., & Rothermel, K. (2014). A classification of location privacy attacks and approaches. *Pers Personal and Ubiquitous Computing*, 18(01), 163–175.
- [23] Jiang, Nanlan, Sai Yang, and Pingping Xu. "Enabling Location Privacy Preservation in MANETs Based on Distance, Angle, and Spatial Cloaking." *Electronics* 9.3 (2020): 458.
- [24] J.-H. Song, V. W. S. Wong, and V. C. M. Leung, "Wireless location privacy protection in vehicular Ad-Hoc networks," *Mobile Networks and Applications*, vol. 15, no. 1, pp. 160–171, 2010.
- [25] Grissa, Mohamed, Attila Altay Yavuz, and Bechir Hamdaoui. "Location privacy in cognitive radios with multi-server private information retrieval." *IEEE Transactions on Cognitive Communications and Networking* 5.4 (2019): 949-962.
- [26] Wu, Zongda, et al. "Constructing dummy query sequences to protect location privacy and query privacy in location-based services." *World Wide Web* 24.1 (2021): 25-49.
- [27] SNAP website, (2018), available: <https://snap.stanford.edu/data/loc-brightkite.html>. (accessed on 9 Oct, 2021).
- [28] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "A Survey on Privacy of Location-Based Services: Classification, Inference Attacks, and Challenges." *Journal of Theoretical & Applied Information Technology* 95.24 (2017).
- [29] Bettini, Claudio. "Privacy protection in location-based services: a survey." *Handbook of Mobile Data Privacy*. Springer, Cham, 2018. 73-96.
- [30] Gupta, Ruchika, and Udai Pratap Rao. "An exploration to location based service and its privacy preserving techniques: a survey." *Wireless Personal Communications* 96.2 (2017): 1973-2007.
- [31] Tefera, Mulugeta K., Xiaolong Yang, and Qifu Tyler Sun. "A Survey of System Architectures, Privacy Preservation, and Main Research Challenges on Location-Based Services." *KSII Transactions on Internet & Information Systems* 13.6 (2019).
- [32] Rajashekar, M. B., and S. Meenakshi Sundaram. "A Survey on User's Location Detail Privacy-Preserving Models." *SN Computer Science* 1 (2020): 1-6.
- [33] Alrahhal, Hosam, et al. "A Symbiotic Relationship Based Leader Approach for Privacy Protection in Location Based Services." *ISPRS International Journal of Geo-Information* 9.6 (2020): 408.
- [34] Mohamad Shady Alrahhal, Maher Khemakhem and Kamal Jambi, "Agent-Based System for Efficient kNN Query Processing with Comprehensive Privacy Protection" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 9(1), 2018. <http://dx.doi.org/10.14569/IJACSA.2018.090108>.
- [35] Mohamad Shady Alrahhal, Muhammad Usman Ashraf, Adnan Abesen and Sabah Arif, "AES-Route Server Model for Location based Services in Road Networks" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 8(8), 2017. <http://dx.doi.org/10.14569/IJACSA.2017.080847>.
- [36] Alrahhal, Mohamad Shady, Maher Khemakhem, and Kamal Jambi. "Achieving load balancing between privacy protection level and power consumption in location based services." (2018).
- [37] Alrahhal, Mohamad Shady, and Adnan Abi Sen. "Data mining, big data, and artificial intelligence: An overview, challenges, and research questions." (2018).
- [38] Mona Alfifi, Mohamad Shady Alrahhal, Samir Bataineh and Mohammad Mezher, "Enhanced Artificial Intelligence System for Diagnosing and Predicting Breast Cancer using Deep Learning" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 11(7), 2020. <http://dx.doi.org/10.14569/IJACSA.2020.0110763>.

# A Conceptual User Experience Evaluation Model on Online Systems

Norhanisha Yusof<sup>1</sup>, Nor Laily Hashim<sup>2</sup>, Azham Hussain<sup>3</sup>

Department of Information Technology and Communication, Politeknik Balik Pulau, Penang, Malaysia<sup>1</sup>  
School of Computing, Universiti Utara Malaysia, Kedah, Malaysia<sup>2,3</sup>

**Abstract**—An online system has become a priority for organisations or companies in many countries, as it allows many processes to be conducted via online platforms, which contributes to profit gain. There are different types of user experience (UX) evaluation models that have been proposed to guide the measurement and development process. However, most of these models only have dimensions, and there is no guidance for UX measurement on online systems. The lack of evaluation models for online system measurement requires further investigation. This paper aims to identify the gaps in UX evaluation models, and develop a conceptual UX evaluation model for online systems. The method used in this study includes reviewing several literatures and shortlisting the relevant publications on UX and online systems. After that, the gaps were identified from the existing UX evaluation model in the relevant publications based on the ISO standard. Then, the study identified the important components of UX, and proposed a new conceptual UX evaluation model for online systems. The results of the study are the identification of the gaps in existing UX evaluation models, and the development of a new conceptual UX evaluation model that is specifically for online systems. Therefore, the results help in considering UX dimensions, criteria, and metrics and potential UX components for evaluation and measurement. The paper contributes to system developers, designers, and also researchers for future UX evaluation model development for online systems. Future studies could use the reviewed UX evaluation models to identify relevant dimensions of online systems, and hence improve the model that they will develop. The findings may also be beneficial to organisations that own online systems by providing guidelines on important dimensions involved in their UX-based evaluations.

**Keywords**—User experience; UX evaluation; UX model; online system; conceptual model

## I. INTRODUCTION

Since the beginning of the millennium, computers have been widely utilised in various countries, resulting in a revolution in information and communication technology (ICT), known as the digital revolution. The digital revolution has rapidly grown in developed countries, where there are various applications or electronic systems in use. Nowadays, an online system is one of the most important applications of ICT that has become a priority in all systems around the world. An online system is a platform used by any individual or organisation to perform their work through the Internet, whereby it is flexible and accessible to users [1]. Online systems might be totally secure, or could allow third-party programmes to join via programming platforms. Online systems are useful tools for storing, organising, utilising, and

gathering data. Nowadays, online platforms have been modernised, whereby the users can have access to all information at their fingertips. The users use the Internet in their daily lives to get information, shop, and communicate with chosen online systems. This priority is quickly expanding in emerging countries because it can help manage any transaction easily. Many governments and businesses support the delivery of online systems, which allows them to leverage existing technology transactions and interactions with higher efficiency and easy access [2]. There are many online systems; for example, eBay, Craigslist, Amazon Marketplace, Airbnb, and Uber [3], [4]. An online system is not only for business purposes, but it is also for government systems, such as e-government [5] and e-procurement system, because it can integrate the use of ICT to improve customer, supplier, and other relationships [6].

The study of online systems is a significant part of today's economy. For example, the potential of an online system to contribute total profitability to governments and companies is enormous. It is important to remember that every dollar saved in terms of cost because of using an online system can directly contribute to the productivity of the country. The study of online systems is significant for a variety of reasons, including the fact that many industrialised countries and business companies use them to manage relationships with their users [4]. The benefits of online systems are as follows: easy to use, quick response time, automatically processed via technology, and faster in terms of accessing information because it is online worldwide.

User experience (UX) is the perception and reaction of real-world users towards real products and services. UX, as defined by ISO 9241-11 (2018), which is the perceptions and responses of a person as a result of the use or anticipated use of a product, system, or service [7]. Besides that, UX is also related to users' emotions, beliefs, and physical and psychological responses. It also incorporates brand image, presentation, system performance, and physical state as a result of earlier experiences, attitudes, skills, and personality, among other things [8]. The term UX is most commonly associated with the design and presentation of online software solutions such as websites and applications [9]. Thus, UX is an important factor in creating quality products, systems, and services, especially for online systems because many users use online platforms in their daily lives and the impact needs to be known through an evaluation. UX evaluation is one way of analysing individuals' experiences [10]. UX evaluation is a burgeoning field with a wide range of approaches [11]. One of the most important

aspects of studying UX is evaluation, which refers to the use of a set of procedures and tools with the goal of determining how people feel about using a system or product [8].

However, there is a lack of investigation of the user experience related to an online system in literature as explored by the authors. For example, the study by [12] explored the emotions of online learning systems among college teachers, and proposed an integrated model of Technology Acceptance Model (TAM); however, they emphasised on identification of technology usage intentions among the teachers, and the study did not involve UX dimensions except emotion only. There are also some researchers who focus more on factors influencing UX [13] but overlooked UX evaluation, including coming up with a conceptual UX evaluation model for an online system with important measurements. Besides that, the study by [14] proposed a conceptual model for UX whereby it is related to the process of UX practices in organisations. The model does not focus on evaluating UX of online systems and does not discuss the model in terms of UX components from instrumental and non-instrumental qualities perspectives. UX components can be defined as important aspects of human-technology interaction [15]. For example, instrumental and non-instrumental qualities are important in a UX model measurement because they are core components of experiences related to the user perception including emotion while users are using the system [16]. On the other hand, [17] have developed a new conceptual UX model for evaluation that focuses on workers' performance evaluation, and not on the online system. Furthermore, the need to conduct this study is motivated by the lack of existing models that were more focused on specific systems such as e-commerce [18], e-banking [19], and mobile [20]. Furthermore, these studies do not provide clear and comprehensive measurement for general online systems.

The gaps discussed above have motivated the authors to study UX that is related to online systems by identifying the required dimensions and main components of the conceptual model that can be referenced for UX measurement on online systems. Therefore, the gaps in existing UX evaluation models need to be identified with a conceptual UX evaluation model development for a better measurement.

The objective of this current research is to identify the gaps in the existing UX evaluation models and develop a conceptual UX evaluation model for online systems. The research questions (RQ) addressed in this paper is as follows:

RQ 1: What are the gaps in UX evaluation model for online systems?

RQ 2: What are dimensions of a conceptual UX evaluation model for online systems?

This paper is organised as follows: Section 2 presents the background of online systems and relates to a user UX; Section 3 discusses the methodology of the study; Section 4 discusses the results and discussion regarding existing UX evaluation model including conceptual model development; and Section 5 provides the concluding remarks and future work for this paper.

## II. BACKGROUND

This section presents the research studies in the literature related to UX evaluation, which covers the content. The gaps in the existing UX evaluation models discovered in the literature review are considered in the newly developed conceptual UX evaluation model for online systems.

### A. Evaluation

UX has become a more prominent part of system development, following the growth of business and process models. UX is influenced by the user's internal state, for example, individual motivations, expectations, needs, and mood; the characteristics of the system being used, such as complexity, usability, and functionality; and it is also associated with interactions of the context or environment with the system [21]. The analysis of user interaction related to web systems is crucial for satisfaction, and it may even encourage changes to improve UX level [22]. There are five points in UX concepts, which are understanding of research, sketch, design, implement, and evaluation [1]. One of the pillars of academic UX research has been identified as evaluation [11]. The experience evaluation is crucial for UX practitioners in the workplace [11]. UX evaluation depends on the components or factors that exist in the models [23]. Thus, UX encompasses product aspects and individual aspects, whereby these consist of users' perception such as pragmatic and hedonic quality, aesthetics, user's emotional experiences, expectations, and needs [21].

In the literature review, many UX evaluation studies of online systems only have dimensions and criteria [15], [18]–[20], [24]. However, they are lacking the UX dimensions, criteria, and metrics in the UX evaluation models. Having the dimension, criteria, and metric of the model for measurement is important because the organisation or stakeholder can have a clear understanding of how to conduct an evaluation of the product or system usage more specifically. Furthermore, there is a need for conceptual UX evaluation models for online systems because this model fills the gap in the literature of UX and online systems, including providing important dimensions; and it is also due to the studies on UX evaluation that did not define additional concrete UX dimensions [11]. Moreover, the conceptual UX evaluation model was developed since there was a lack of user experience studies on online systems after the publication review was conducted by the researcher. From the reviews conducted on this domain and to the best of the authors' knowledge, there is no work focused on the UX evaluation model for online systems yet. Furthermore, UX is still a blurred conceptualisation, causing the need for this study to be conducted [9]. Therefore, this study has developed a conceptual UX evaluation model for online systems.

### B. UX of Online Systems

With the growth of Information and Communication Technology (ICT) nowadays, information can be obtained through an online system, social media, chat, and others [25]. Through the circulation of technology nowadays, many organisations or companies use online system platforms for various processes and transactions. It is publicly accessible, open, and more convenient. Thus, the user experience should be designed for human use as well, with easy access and

assistance for user limitations such as reading small texts [1]. However, there is still a lack of research conducted related to UX and evaluation for online systems. From a conceptual standpoint, different researchers have varying interpretations of what UX is to suit their studies and application needs. Some studies argue that UX is holistic; others suggest that the complexity of experience should be broken down into evaluative components.

Based on the analysis of the existing models, it has been noted that the field of UX research should also be given attention in terms of measurement models [13]. UX can be associated with users' affective, cognitive views such as hedonic quality, attractiveness, and the subjective perception made by the users of a product, system, or services [9]. However, additional research is needed into how these qualities with UX dimensions can be represented in a more precise manner. Moreover, certain important dimensions that have a significant impact on user experience remain implicit in certain models [13]. Thus, a conceptual model for UX modelling should be established to comprehend, explore, and analyse interactions between users and products [13]. Defining a conceptual UX evaluation model can be considered by incorporating UX dimensions, criteria, and metrics in modelling and reflecting elements that have a direct impact on the user experience and online systems. The author in [17] developed a new conceptual UX model for evaluation called the TAMUX model, but it is for workers' performance evaluation and not for online systems. TAMUX is the combination of the Technology Acceptance Model (TAM) and the Components of user experience (CUE) model. The weaknesses of the conceptual model by [17] is it is lacking in terms of dimensions with criteria because it only has components such as task characteristic, individual characteristic, system characteristic, and performance effect.

Meanwhile, the current study by [25] conducted an evaluation of an online learning programme to investigate user experience improvement. However, the evaluation in Chow's research is more concerned about usability with the five dimensions of learnability, efficiency, memorability, errors, and satisfaction. There is no conceptual UX model that can be used or referred to for online system measurement. Therefore, these studies show the need for a conceptual model that consists of UX dimensions, criteria, and metrics because the measurement can be conducted in a more clearly and directed manner. Moreover, the development of conceptual the UX evaluation model also helps or enables organisation or business stakeholders to define and redefine the metrics in order to improve their user interface and users' experience [26]. The next section will discuss further on methodology and analysis of findings to answer the RQ as stated in the earlier section.

### III. METHODOLOGY

This study used online databases such as 'Scopus' and 'Elsevier Science Direct' to search for previous literature. The searching technique in this study is referring to [27] study as a guidance. The review considers publications on user experience, UX, models, and systems for nine years from January 2013 to 15 October 2021. Only articles written in English and final published journals were considered for this

study. Keywords such as 'user experience', 'UX', 'model', and 'online system' were used to search for articles in databases. 'Paper title' or 'abstract' or 'paper keywords' were the search criteria. The researcher found an entire list of all relevant publications by using multiple keyword combinations. From all these databases, the initial search yielded roughly 503 research publications.

Then, 503 of these papers were screened based on title and abstract. The screening method eliminated articles such as reports, book reviews, and review papers. The papers that passed the screening were subsequently scrutinised for appropriateness for online systems by evaluating a manuscript. Approximately 227 papers were screened through manuscript review and selection criteria. Then, only six papers were shortlisted and further investigated after reading the entire text. After completing the reading, irrelevant publications were filtered out based on the scholarly judgement. The result of final papers shows that there is a lack of UX model evaluation for online systems. Table I shows the results of the final review from databases, which resulted in a final shortlist of six relevant publications by the authors. While Fig. 1 shows the flow of literature review method.

TABLE I. REVIEW PAPERS FROM DATABASES

| Databases               | Keyword used                                      | Result based on keyword | Result based on title and abstract | Final shortlist |
|-------------------------|---------------------------------------------------|-------------------------|------------------------------------|-----------------|
| Scopus                  | 'user experience' AND 'model' AND 'online system' | 105                     | 24                                 | 4               |
|                         | 'UX' AND 'model' AND 'online system'              | 3                       | 0                                  | 0               |
| Elsevier Science Direct | 'user experience' AND 'model' AND 'online system' | 337                     | 192                                | 2               |
|                         | 'UX' AND 'model' AND 'online system'              | 60                      | 11                                 | 0               |
| Total                   |                                                   | 503                     | 227                                | 6               |

After conducting a final review of relevant publications related to online systems, the authors analysed the papers in terms of domain of UX Model, instrument used, number of participants, and dimensional measurement, which are based on System and software Quality Requirements and Evaluation (SQuARE) such as Measurement of quality in use (ISO 25022:2016) and Measurement of system and software product quality (ISO 25023:2016). These two ISOs were used as a guidance to determine the dimensions that have similar descriptions in the various existing UX evaluation models, and to ensure consistency of the terms used in the identification of common dimensions in this domain. ISO standards are widely used in the field of Human-Computer Interaction (HCI) [28]. The current study has reviewed six papers that were shortlisted in order to identify the gaps in the existing UX evaluation models and to answer research question 1 (RQ 1). Meanwhile, the components of a conceptual UX evaluation model for

online systems have been developed based on the gaps identified and included the important dimensions from the literature in order to answer the RQ 2. Therefore, the study answered the RQ by conducting the paper review, identified the gaps, and then developed a new conceptual UX evaluation model.

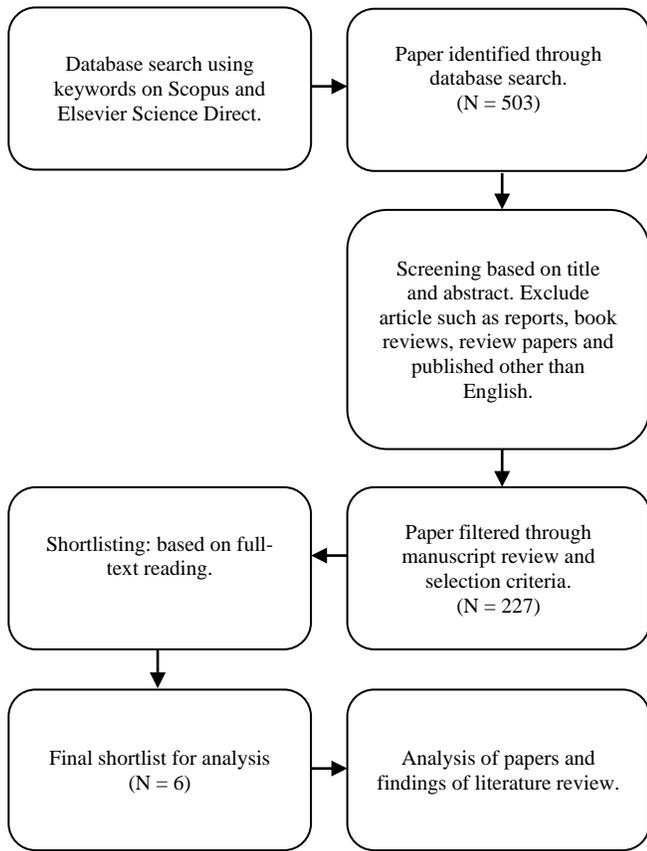


Fig. 1. Flow of Literature Review Methodology.

#### IV. RESULT AND DISCUSSION

There are various UX evaluation models or frameworks in the literature that have been introduced. However, these models are inadequate for online system measurement because these models are lacking in terms of important dimensions, criteria, and metrics. This study used ISO 25022:2016 and ISO 25023:2016 as guidance by referring to their description and categorisation of the dimensions in the UX evaluation model since there are various dimensional measurements in the literature. This section discusses existing UX evaluation models in the literature that have been identified in Section 3, and a summary of these models is presented in Table I by extracting important aspects such as domain of UX Model, instrument used, number of participants, dimensional measurement, other dimensions used (Based on ISO 25022:2016 and ISO 25023:2016), and the identified remarks or gaps.

##### A. E-commerce’s UX Evaluation Model

In the work by [18], it is mentioned that improved user experience can attract new users, increase visitor numbers, and transaction volume of the system. The model was developed to

evaluate e-commerce websites. The study’s strength is that it employs the Fuzzy Comprehensive Evaluation Method, which improves the model by implementing specific steps to enable systematic evaluation, including the use of triangular fuzzy number theory, methods of psychological stratification, and weighting factors. Six experts in the field of e-commerce evaluated the model. The model developed by [18] consists of three dimensions: Visceral, behavioural, and reflective, and has verification. Based on ISO 25022:2016 and ISO 25023:2016, this model has emotional and functionality dimensions. However, this model does not provide dimension with criteria, and no metric is provided for online system measurement. Besides that, among the weaknesses of the model is that information can be lost by criteria’s aggregation as it is based on probability and possibility measurement [through Analytic Hierarchy Process (AHP)] [29], [30]. Therefore, there is a need to provide dimensions, criteria, and metrics of the proposed model for online systems because these three components will provide clear measurement for evaluation. Fig. 2 shows the model developed.

##### B. Testing of UX Model with News Sites

The study by [15] has established a model that constitutes a fundamental theory for evaluating news websites, and is theoretically applicable to all UX studies. The model is expanded to estimate how satisfied users are with news websites. Instrumental qualities, non-instrumental qualities, and emotional responses are among the UX components that exist in this model. The strength of the model is determined by the measurement of instrumental qualities and non-instrumental qualities using AttrakDiff2 [31]. Meanwhile, emotional responses are measured through the Positive and Negative Affect Schedule (PANAS) elements [32]. Based on the ISO standard, this model consists of usefulness, emotional (ISO 25022:2016), and trust dimensions (ISO 25023:2016). However, the weaknesses of this model are broad measurements. For instance, pragmatic and hedonic qualities do not have criteria and metrics that are specifically for evaluation measurement. The model is also only for testing news sites and online news. Therefore, there is a need to provide dimensions, criteria, and metrics of the proposed model for online systems. Fig. 3 shows the developed model for testing news sites.

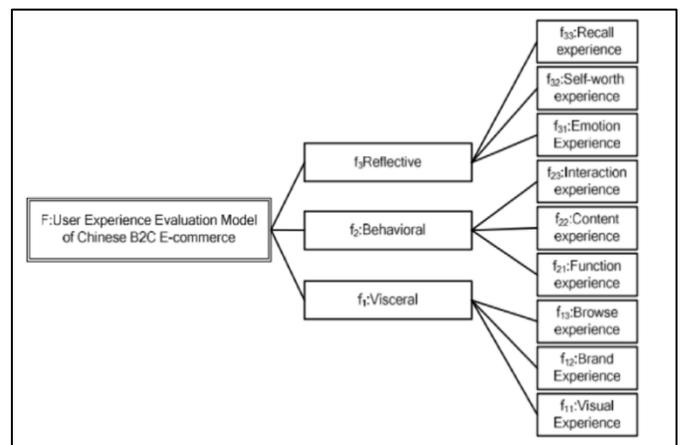


Fig. 2. Improved Hierarchical Model of p E-commerce’s [18].

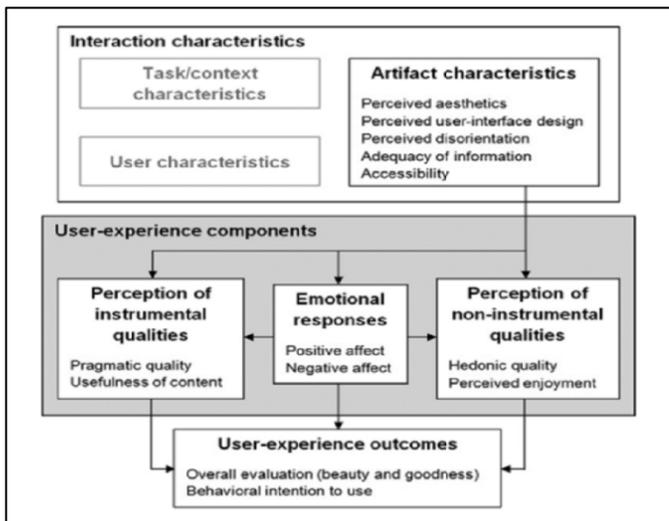


Fig. 3. High-Level Representation of the Model of UX with News Sites [15].

### C. Security and Usability Evaluation Model for e-Banking

The work by [19] investigated frameworks or models, including approaches for evaluating e-banking security and usability. Their study combines a collection of frameworks that were related to security and usability properties such as the following: (1) interface; (2) navigation; (3) content; (4) offered services; (5) registration and transaction procedure; and (6) multi-factor authentication methods. This combination is called the hybrid security and usability evaluation model. Based on the ISO 25023:2016 standard, this model consists of usability and security dimensions. The strength of this model is it is able to evaluate e-bank assets that are accessible to the public, and it covers 13 different security and usability categories with more than 160 metrics. Finally, this model comprises the security evaluation part of the framework with 72 metrics, and usability evaluation part with 97 metrics. However, the weakness of this model is the unclear presentation of dimensions in the model with criteria and metrics. Moreover, security and usability evaluations are required to be considered in the quality improvement process; thus, these dimensions need to be designed and tested as part of the quality improvement process in order to ensure their coherence with other parts of the process [19]. Therefore, it can be concluded that the presentation of this model can be improved for evaluation and online system measurement. Fig. 4 shows the proposed security and usability evaluation model.

for the edutainment field. Moreover, the model can be used in a variety of disciplines, including industrial and collaborative settings. Based on the ISO standard, this model consists of emotional (ISO 25022:2016) and usability dimensions (ISO 25023:2016). However, the weaknesses of this model are this model is general due to the lack of criteria and metrics that link to the dimensions. The measurement for the online system should be clear in order to get better results for evaluation. Therefore, the dimensions, criteria, and metrics of the proposed model for online systems are required. The UXIVE model is depicted in Fig. 5.

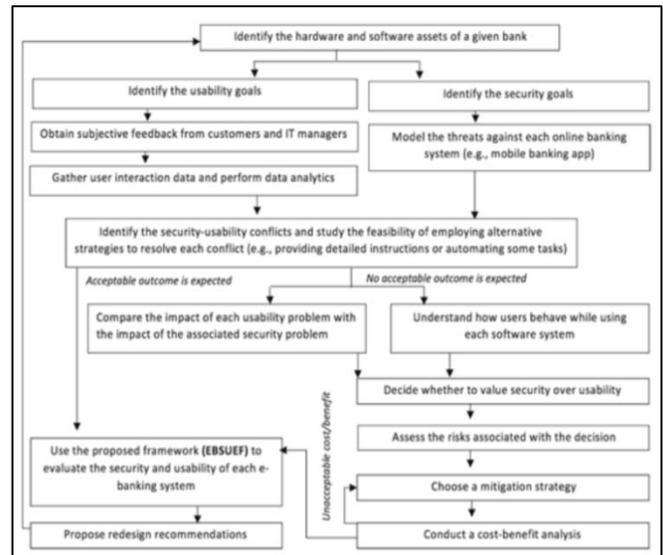


Fig. 4. The Proposed Security and usability Evaluation Model [19].

### D. Evaluation of the Modified Immersive Virtual Environment (UXIVE) Model

The UXIVE model has been introduced by [24], in which the constructed model components are derived from existing models. The method used was quantitative (questionnaire), and the findings were analysed for model validation using Structural Equation Modelling (SEM). The sample size is large (152 respondents); thus, SEM has been employed for statistical analysis. The strength of this model is validated, which comprises ten new UX dimensions, namely presence, immersion, engagement, skill, emotion, flow, usability, technology adoption, judgement, and experience consequence

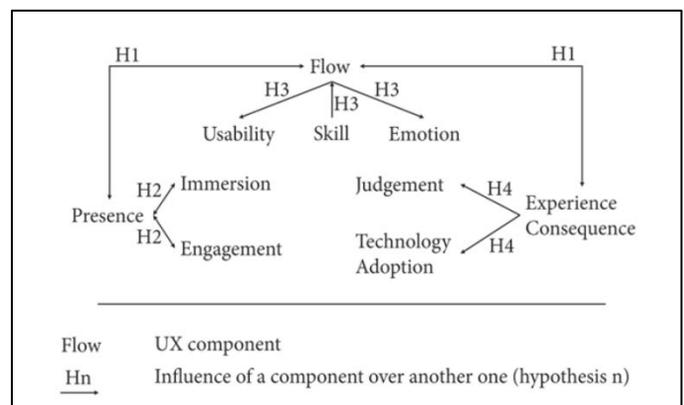


Fig. 5. The Hypothesised UXIVE Model [24].

### E. UX Evaluation Model on Mobile Terminal Products

The Content Interaction Vision Model (CIV Model) was developed by [20]. Content, interaction, and vision are the three UX dimensions of the CIV model. Their study examines the important dimensions by using the Heuristic evaluation approach and Nielsen's ten (10) usability principles. The strength of the model is that the results of the dimensions are displayed in the form of a radar chart, which uses a set of questionnaires from the System Usability Scale (SUS) for the evaluation. Based on ISO 25023:2016, this model consists of functionality and usability dimensions. In the same vein, the

study by [33] also used SUS and proposed the UX Maturity Model for e-commerce websites, which focused on the three cores of user experience, namely user research, visual design, and user testing. This shows that the UX evaluation conducted with other instruments could strengthen the findings of the study. However, Huang *et al.*'s [20] model evaluation focused more on the usability principle in order to construct the CIV evaluation method and it has a subjective metric. Moreover, this model has no verification for the feasibility of the evaluation model. Therefore, it is necessary to identify the appropriate dimension, criteria, and metrics for online systems to get clear measurement of evaluation regarding user experience (UX). Fig. 6 shows the CIV Evaluation Model of User Experience.

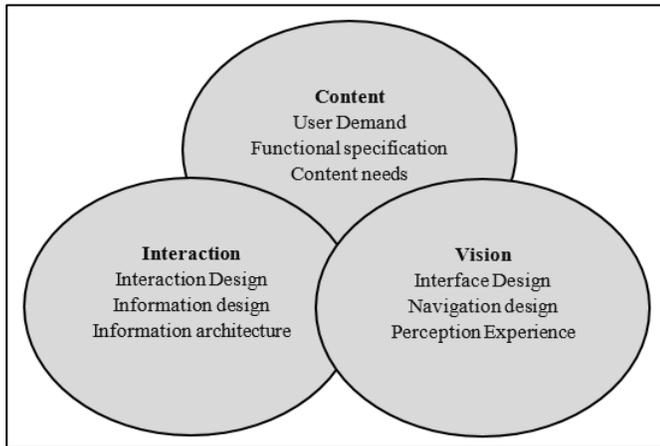


Fig. 6. CIV Evaluation Model of user Experience [20].

#### F. Loyalty Model for E-commerce Recommender Systems

The study by [34] about the use of an e-commerce recommender agent explores the major determinants in the establishment of female online shopper loyalty. Then, a new model is introduced, developed, and analysed in order to improve e-commerce consumer loyalty via the recommender systems. The strength of the model is the use of SEM to analyse the relationships between independent and dependent variables for quantitative research. Their study used the SEM tool because it combines factor analysis, path analysis, and multiple regression analysis to evaluate construct relationships [34]. Moreover, based on ISO 25023:2016, the dimensions in this model that were measured are usability, transparency, satisfaction, and trust. However, this model did not provide clear criteria and metrics for evaluation, which consists of dimensions and relationships for each construct. Their research also has some limitations in terms of data dissemination and collection because the evaluation only involved one e-commerce platform, whereby it could not be generalised to the population studied. In addition, the proposed model requires two items: Experience and search-characteristic products for

future research [34]. Therefore, there is a need for dimensions with criteria, and metrics of the model as well as a need to consider the experience element of the model in future research. Fig. 7 shows the Loyalty Model for E-commerce Recommender Systems. Meanwhile, Table II shows the summary comparison of existing UX evaluation models or frameworks identified.

Based on the model analysis in Table II above, the models were purposely developed for e-commerce [18], testing a news site [15], e-banking system [19], edutainment [24], mobile terminal products [20], and an e-commerce recommender system [34]. Thus, there is no generic online system measurement that can be used as each one is tailored to a specific system. This means that existing UX models are more focused on systems such as e-commerce, e-news, e-banking, and mobile product but there is no guided measurement for online system as overall such as for online reservation, online booking including e-government, e-procurement.

The model by [15] has provided subjective metrics, but it is for news websites. Meanwhile, Huang *et al.* [20] did not present how they performed verification for the feasibility of the evaluation model. In addition, the [34] model does not have criteria for the measurement. Moreover, the model by Liu *et al.* [18], Alarifi *et al.* [19], and Tcha-Tokey *et al.* [24] also did not provide metrics for UX evaluation. The dimensions with criteria and metrics in the existing model are lacking in order to give clear description to the users about the measurement and conducting evaluation. Thus, these studies show that there is a lack in terms of dimensions with criteria and metrics in the existing models that are specifically for online system measurement. Hence, this current study motivates the researchers to develop a new conceptual UX evaluation model for online systems by identifying the important dimensions for better measurement.

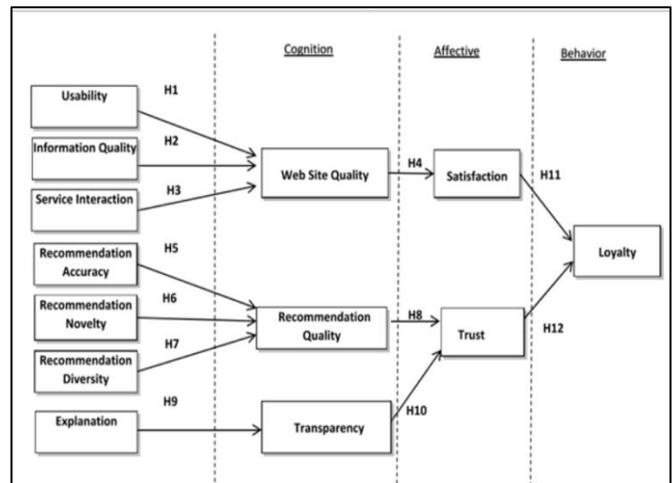


Fig. 7. Loyalty Model for E-commerce Recommender Systems [34].

TABLE II. THE SUMMARY COMPARISON OF EXISTING UX EVALUATION MODELS OR FRAMEWORKS

| Authors                       | Instrument and Number of Participants       | Dimensions that measured                                                                                  | Other Dimensions (Based on ISO 25022: 2016 and ISO 25023: 2016) | Remarks/ Gaps                                                                                                                    |
|-------------------------------|---------------------------------------------|-----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|
| Liu <i>et al.</i> [18]        | Questionnaire (6 experts)                   | Visceral, Behavioural, Reflective.                                                                        | Functionality, Emotional.                                       | No UX metric provided in the model.<br>Model for e-commerce site.                                                                |
| Aranyi and van Schaik [15]    | Questionnaire AttrakDiff2, PANAS (85)       | Ease of use, Perceived Enjoyment, Beauty and Goodness, Behavioural intention.                             | Usefulness, Trust, Emotional, Satisfaction.                     | Has subjective metrics but it is for news websites.<br>Model for designer (online news or interactive product)                   |
| Alarifi <i>et al.</i> [19]    | Users using web portals from the five banks | -                                                                                                         | Usability, Security                                             | Do not provide criteria and metrics of model.<br>Does not provide any prioritisation of metrics.<br>Model for e-banking systems. |
| Tcha-Tokey <i>et al.</i> [24] | Questionnaire (152)                         | Presence, engagement, immersion, flow, skill, experience consequence, judgement, and technology adoption. | Usability, Emotional                                            | Only have UX dimensions.<br>No metrics provided in the model.<br>Model for edutainment field.                                    |
| Huang <i>et al.</i> [20]      | Experiment (10)                             | Content Interaction Vision                                                                                | Functionality, Usability                                        | No verification for the feasibility of the evaluation model.<br>For mobile and design of online product only                     |
| Ali <i>et al.</i> [34]        | Survey questionnaire (300)                  | Web Site Quality, Recommendation Quality, Loyalty                                                         | Usability, Transparency, Satisfaction, Trust                    | Has dimensions whereby it is a construct, but do not have criteria.<br>Model for e-Commerce Recommender Systems.                 |

Moreover, it also shows that there are various dimensions measured by the existing models, and the researcher used the ISO standard such as System and Software Quality Requirements and Evaluation (SQuaRE) which Measurement of quality in use (ISO 25022:2016) and Measurement of system and software product quality (ISO 25023:2016) as a guidance in order to identify and categorise appropriate UX dimensions that are similar to each model. Thus, the models by Alarifi *et al.* [19], Tcha-Tokey *et al.* [24], Huang *et al.* [20], and Ali *et al.* [34] have similar descriptions of dimensions, for example, usability. Meanwhile, the model by Liu *et al.* [18] and Huang *et al.* [20] have the functionality dimension. Besides that, the emotional dimension has been measured by the model by Liu *et al.* [18], Aranyi and van Schaik [15], and Tcha-Tokey *et al.* [24]. Furthermore, the model by Aranyi and van Schaik [15] and Ali *et al.* [34] have the trust and satisfaction dimension in their measurement. Therefore, this argument shows that the common UX dimensions for evaluation for online system measurement are usability, functionality, emotional, trust, and satisfaction. These dimensions can be considered in measurement for online systems.

On the other hand, based on ISO 25022:2016, the usability characteristic comprises efficiency, effectiveness, and satisfaction, whereby these dimensions can be considered in the model for online system measurement. However, from the analysis that has been done, the measurement of these

dimensions requires UX components, so that the evaluation of users can be conducted more accurately and correctly, for example, considering the pragmatic and hedonic qualities with UX metrics. Based on Fig. 8 below, it shows the gaps flow from existing UX evaluation models that were derived from Table II. The purpose of this figure is to provide an overview of the models that have been identified from the literature, including the gaps and the requirements of the developed conceptual UX evaluation model. As mentioned earlier, this study identified six existing evaluation models related to user experience and online systems. The flow in Fig. 8 shows that the UX evaluation model requires UX dimensions, criteria, and metrics for the model of online systems in order to get a clear measurement illustration.

Besides that, the findings also revealed that any development of the UX evaluation model for online systems should consider prioritisation for the metrics because it will determine which dimensions need to be prioritised for the evaluation measurement [19]. Furthermore, there is a need for the verification for the feasibility of the model to be concerned in any UX evaluation model development. The existing UX model is also still not focus on measuring emotions. Therefore, it can be concluded that there is a need for dimensions with criteria and metrics for online system measurement. Among the dimensions that can be considered in the measurement are efficiency, effectiveness, satisfaction, functionality, emotional, and trust. Thus, these findings addressed important components with justification for an online system measurement.

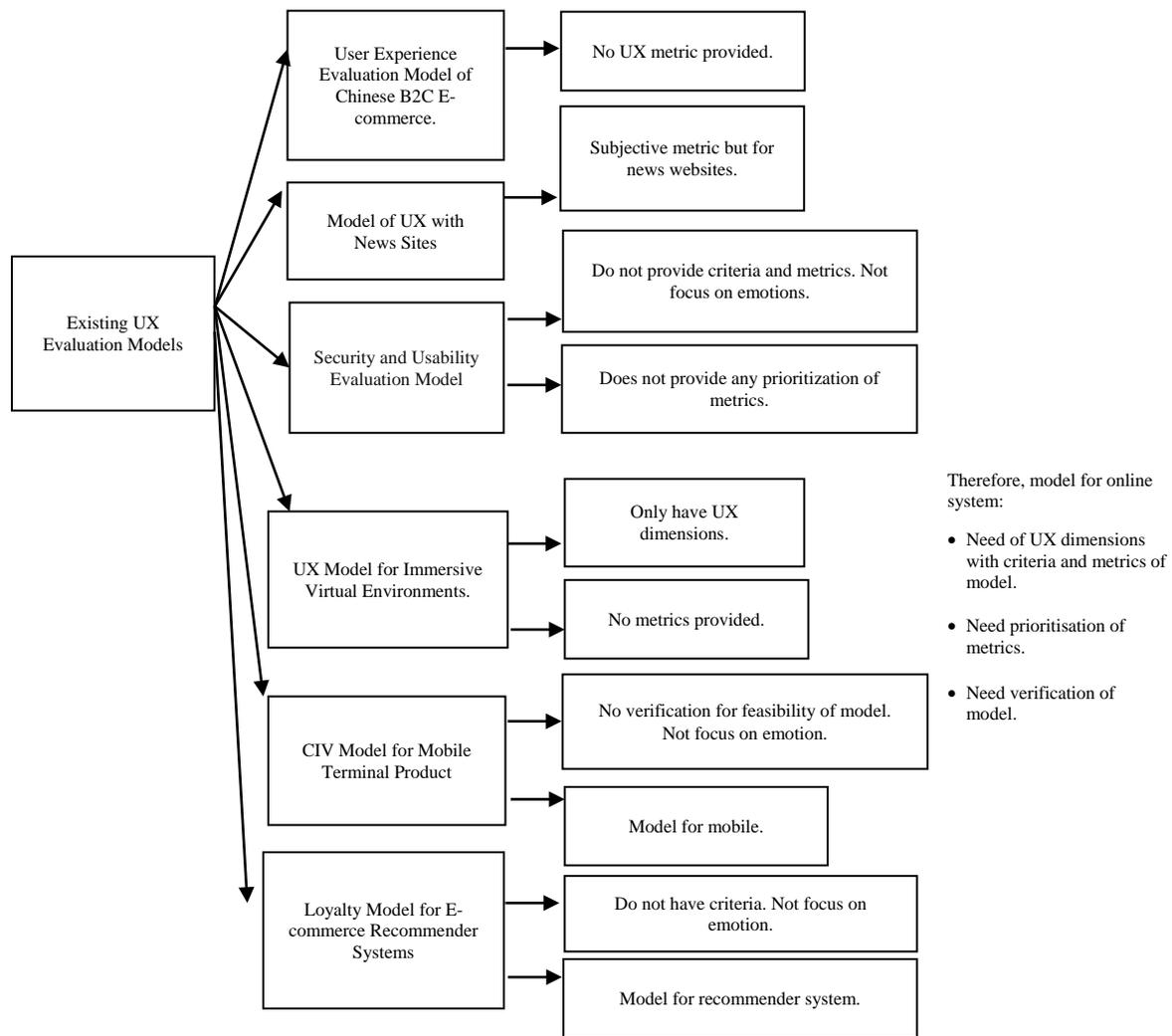


Fig. 8. Gaps of Existing UX Evaluation Models.

### G. Development of Conceptual UX Evaluation Model for Online Systems

Fig. 9 shows the conceptual UX evaluation model for online systems. This conceptual model was proposed to fill the gaps in the literature that has been identified and discussed in Section 4. The conceptual model in this study is adapted from the Aranyi and van Schaik model [15] because this model consists of interaction characteristics such as user characteristic, task characteristic, and artifact characteristic; and user experience components such as perception of instrumental qualities, perception of non-instrumental qualities, and emotional responses. As mentioned earlier, all these components are important and needed because they are core components of experience whereby users' perception, including emotions, are evaluated during the interaction with the system [16]. User characteristic refers to user knowledge or skill such as education, system experience, personality or role, and language proficiency [35]. While task characteristic refers to business workflow in the study by Seffah *et al.* [36], another study refers to complexity and involvement in primary tasks [35]. Besides that, artifact characteristic refers to the online system used by the user.

There are many examples of online systems such as online ticketing systems, online management systems, online billing systems and so on. Online systems are also implemented by the e-government, e-procurement systems, and others. Thus, the findings are significant for these type of online systems to be considered because they can refer to this conceptual UX evaluation model that will be developed for system measurement and enhance their positive user experiences. Meanwhile, instrumental qualities can be related to technical features, for example, task suitability, self-descriptiveness, and controllability [16]. Non-instrumental qualities can be related to design features, for example, material, form, and combinations of colour [16]. The perception of non-instrumental qualities, emotions, and the perception of instrumental qualities are influenced by interaction characteristics, which consist of system function, user, and context, whereby it has experiential consequences such as overall experience, acceptance, intention to use, and alternative choice [37]. Thus, these core UX components of experience are important for measurement by the organisation, especially those that use online systems.

Based on the findings in Table II which have been discussed, a conceptual model needs to have dimensions with criteria and metrics of model, prioritization of metrics, and model verification. Moreover, based on the discussion earlier (Table II), the dimensions that can be considered in the measurement are efficiency, effectiveness, satisfaction, functionality, emotional, and trust. However, the conceptual model development is designed by including appropriate dimensions and linking with criteria and metrics in order to provide clear measurement for online systems. Another study has provided UX dimensions with metrics, but it is for mobile learning [38].

According to [39], instrumental qualities are related to efficiency, effectiveness, navigation, system visibility, and others. Meanwhile, [40] discussed that non-instrumental qualities are related to aesthetics, innovativeness, and originality. There has been research that places satisfaction under the category of instrumental qualities [39]. However, attractiveness and satisfaction could be placed under non-instrumental qualities or hedonic qualities based on literature support and data from systematic literature review (SLR) as conducted by [41] study. Based on this argument, satisfaction can be measured as a hedonic quality, whereby it considers the overall experiences of the system by the users [39]. Therefore, based on these arguments, efficiency and effectiveness can be placed in the instrumental qualities, while attractiveness and satisfaction can be placed in the non-instrumental qualities as shown in Fig. 9. These dimensions also show that the proposed conceptual UX evaluation model for online system has been extended from [15] terms of dimensions as shown in Fig. 9.

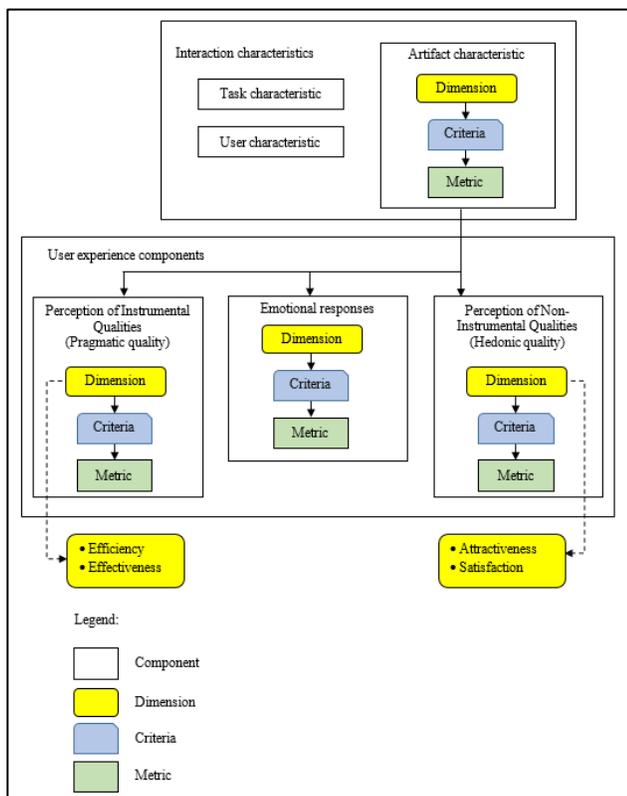


Fig. 9. A Conceptual UX Evaluation Model for Online Systems.

Based on Fig. 9, a conceptual UX evaluation model for online systems has two parts, which are interaction characteristic and user experience components. Interaction characteristics consist of components such as task characteristic, user characteristic, and artifact characteristic. The conceptual model that has artifact characteristic links to user experience components, which has perception of instrumental qualities (pragmatic quality), emotional responses, and perception of non-Instrumental qualities (hedonic quality) whereby it is adapted from Aranyi and van Schaik's model [15]. The new conceptual UX evaluation model is extended by including the dimensions, criteria, and metrics in the user experience components. This conceptual model also revealed the important dimension for online systems, namely efficiency and effectiveness for pragmatic quality; and attractiveness and satisfaction for hedonic quality. On the contrary, the study by [7] stated that the sub-category of emotion relates to attractiveness, enjoyment, and fulfilment. However, this study placed attractiveness under hedonic as investigated by [41] study. Besides, instrumental and non-instrumental qualities could influence the reactions of users emotionally in the use of the system [42].

By applying this conceptual UX evaluation model for online system measurement, the organisation can provide a positive experience for their users with the system. Therefore, this conceptual UX evaluation model for online systems can contribute to system developers and designers as a guidance in order to measure users' experiences because the model representation is flexible, simple, and easy to understand. In addition, the novelty of this model development consists of the UX dimensions, criteria, and metrics whereby it will give a clearer structure of the model for evaluation measurement not only to the system developer and designers, but also benefits researchers for future studies.

## V. CONCLUSION AND FUTURE WORK

Online systems are becoming more important sources of services for many people; research on the model of online systems is a necessity because it is important for any quality improvement process. This paper conducted a literature review from online databases such as Scopus and Elsevier Science Direct about the current state of UX evaluation for online systems, and the researchers have identified relevant papers that were finally selected for full review and critically analysed. It seems that the existing UX evaluation models do provide dimensions, however, they do not have criteria and metrics in supporting for more detailed and guided the measurement. It is found that many of the developed UX evaluation models have weak links between dimensions, criteria, and metrics in the evaluation measurement. The analysis of papers was based on System and Software Quality Requirements and Evaluation (SQuaRE) such as Measurement of quality in use (ISO 25022:2016) and Measurement of system and software product quality (ISO 25023:2016). After identifying gaps in the previous literature on existing UX evaluation models, then this study proposed a new conceptual UX evaluation model for online systems and extended the model by adding important dimensions needed for measurement such as efficiency, effectiveness, attractiveness and satisfaction. Moreover, the user experience component

such as pragmatic quality, emotional responses and hedonic quality also are needed for measurement. The newly developed conceptual UX evaluation model is expected to aid decision makers in resolving any evaluation issues, for instance, the early stages in the development process of the model for online systems. Thus, the goal of this study has been achieved by analysed the UX evaluation models and proposed a conceptual UX evaluation model for online systems that can benefit system developers, designers, and also researchers. It believes that a developed model consisting of dimensions, criteria, and metrics is necessary to ensure the comprehensiveness of the online system measurement in the future.

Future research can be conducted to explore more UX dimensions for online systems in order to get more understanding for evaluation. For future studies, further identification of general criteria and metrics for online systems will be conducted. Therefore, this paper provides a new conceptual UX evaluation model for online systems whereby it is significant to the system developers and designers because they can use the findings of this study in the system development phase and researchers can use it as a guide for future studies. Findings from this paper are also important for system developers and designers to gain an in-depth understanding of the important dimensions of online system measurement such as for e-commerce and it can be used as a basis or guide to redesign existing systems to enhance positive user experience.

#### ACKNOWLEDGMENT

This research was supported by Ministry of Higher Education (MoHE) of Malaysia through Fundamental Research Grant Scheme (FRGS/1/2019/ICT01/UUM/02/4). The authors also want to thank Department of Polytechnic and Community Colleges Education (Hadiah Latihan Persekutuan) for their support.

#### REFERENCES

- [1] J. Sriarunrasmee and C. Anutariya, "The Development of One Stop Service Online System based on User Experience Design and AGILE Method," in ACM International Conference Proceeding Series, 2020, no. January, pp. 64–69.
- [2] P. Sukmasetya, H. B. Santoso, and D. I. Senses, "Current E-Government Public Service on User Experience Perspective in Indonesia," in International Conference on Information Technology Systems and Innovation (ICITSI), 2019, pp. 159–164.
- [3] M. Luca, "Designing online marketplaces: Trust and reputation mechanisms," *Innov. Policy Econ.*, vol. 17, no. 1, pp. 77–93, 2017.
- [4] A. Rangaswamy, N. Moch, C. Felten, G. van Bruggen, J. E. Wieringa, and J. Wirtz, "The Role of Marketing in Digital Business Platforms," *J. Interact. Mark.*, vol. 51, pp. 72–90, 2020.
- [5] M. R. Razlini, "Challenges and issues in Malaysian e-government," *Electron. Gov.*, vol. 13, no. 3, pp. 242–273, 2017.
- [6] K. K. Soong, E. M. Ahmed, and K. S. Tan, "Factors Affecting Malaysia's SMEs in Using Public Electronic Procurement," *J. Inf. Knowl. Manag.*, vol. 19, no. 2, pp. 1–22, 2020.
- [7] L. Hasan, "Examining User Experience of Moodle e-Learning System," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 358–366, 2021.
- [8] D.-O. Ignacio, L. Gustavo, L. Quesada, and L. A. Guerrero, "UX Evaluation with Standardized Questionnaires in Ubiquitous Computing and Ambient Intelligence: A Systematic," *Adv. Human-Computer Interact.*, pp. 1–22, 2021.
- [9] L. Luther, V. Tiberius, and A. Brem, "User experience (UX) in business, management, and psychology: A bibliometric mapping of the current state of research," *Multimodal Technol. Interact.*, vol. 4, no. 18, pp. 1–19, 2020.
- [10] V. Klisman et al., "LogMe: An Application for Generating Logs in Immersive Interactions for UX Evaluation," in Proceedings of the 23rd International Conference on Enterprise Information Systems (ICEIS 2021), 2021, vol. 2, pp. 549–556.
- [11] I. Pettersson, F. Lachner, A. K. Frison, A. Riener, and A. Butz, "A bermuda triangle? - A review of method application and triangulation in user experience evaluation," in Conference on Human Factors in Computing Systems - Proceedings, 2018, pp. 1–16.
- [12] Y. Qu and I. H. Chen, "Are emotions important for college teachers' intentions to use the online learning system? An integrated model of TAM and PAD," *Int. J. Inf. Educ. Technol.*, vol. 11, no. 2, pp. 73–83, 2021.
- [13] B. Yang, Y. Liu, Y. Liang, and M. Tang, "Exploiting user experience from online customer reviews for product design," *Int. J. Inf. Manage.*, vol. 46, no. October 2018, pp. 173–186, 2019.
- [14] C. M. MacDonald, "User Experience (UX) Capacity-Building: A Conceptual Model and Research Agenda," in Proceedings of the 2019 on Designing Interactive Systems Conference, 2019, pp. 187–200.
- [15] G. Aranyi and P. van Schaik, "Testing a model of user-experience with news websites," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 7, pp. 1555–1575, 2016.
- [16] M. Minge and M. Thüring, "Hedonic and pragmatic halo effects at early stages of User Experience," *Int. J. Hum. Comput. Stud.*, vol. 109, no. January 2018, pp. 13–25, 2017.
- [17] M. Mazmela, G. Lasa, E. Aranburu, I. Gonzalez, and D. Reguera, "TAMUX model for industrial HMI evaluation from UX and task performance perspective," *ACM Int. Conf. Proceeding Ser.*, vol. 19, pp. 1–2, 2018.
- [18] X. X. Liu, Q. Y. Liu, W. Wang, C. M. Huang, and W. S. Li, "Research on the E-commerce's user experience evaluation model," *Appl. Mech. Mater.*, vol. 427–429, pp. 2859–2863, 2013.
- [19] A. Alarifi, M. Alsaleh, and N. Alomar, "A model for evaluating the security and usability of e-banking platforms," *Computing*, vol. 99, no. 5, pp. 519–535, 2017.
- [20] Z. Huang, Y. Hong, and X. Xu, "Design and research on evaluation model of user experience on mobile terminal products," in International Conference on Applied Human Factors and Ergonomics, 2020, pp. 198–206.
- [21] G. Pisoni, "Moodle vs Sakai: Evaluating user experience for online entrepreneurship education," in ICETA 2019 - 17th IEEE International Conference on Emerging eLearning Technologies and Applications, Proceedings, 2019, pp. 836–840.
- [22] K. E. S. Souza, M. C. R. Seruffo, H. D. De Mello, D. D. S. Souza, and M. M. B. R. Vellasco, "User Experience Evaluation Using Mouse Tracking and Artificial Intelligence," *IEEE Access*, vol. 7, pp. 96506–96515, 2019.
- [23] I. Atoum, J. Almalki, S. M. Alshahrani, and W. Al Shehri, "Towards Measuring User Experience based on Software Requirements," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 325–331, 2021.
- [24] K. Tcha-Tokey, O. Christmann, E. Loup-Escande, G. Loup, and S. Richir, "Towards a model of user experience in immersive virtual environments," *Adv. Human-Computer Interact.*, vol. 2018, pp. 1–10, 2018.
- [25] L. Chow, "BINUS Online Learning Web User Experience Improvement," *Eng. Math. Comput. Sci. J.*, vol. 2, no. 1, pp. 5–13, 2020.
- [26] I. Kangas, M. Schwoerer, and L. J. Bernardi, "Recommender systems for personalized user experience: Lessons learned at Booking.com," in 15th ACM Conference on Recommender Systems, RecSys 2021, 2021, pp. 583–586.
- [27] S. Nandankar and A. Sachan, "Electronic procurement adoption, usage and performance: A literature review," *J. Sci. Technol. Policy Manag.*, vol. 11, no. 4, pp. 515–535, 2020.
- [28] S. Ashok et al., "Measuring Public Value UX based on ISO / IEC 25010 Quality Attributes," 3rd Int. Conf. User Sci. Eng., pp. 56–61, 2014.

- [29] R. D. F. S. M. Russo and R. Camanho, "Criteria in AHP: A systematic review of literature," *Procedia Comput. Sci.*, vol. 55, pp. 1123–1132, 2015.
- [30] S. Oguztimur, "Why fuzzy analytic hierarchy process approach for transport," *Res. gate*, no. September, pp. 1–19, 2015.
- [31] M. Hassenzahl, S. Diefenbach, and A. Göritz, "Needs, affect, and interactive products - Facets of user experience," *Interact. Comput.*, vol. 22, no. 5, pp. 353–362, 2010.
- [32] D. Watson, L. A. Clark, and A. Tellegen, "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales," *J. Pers. Soc. Psychol.*, vol. 54, no. 6, pp. 1063–1070, 1988.
- [33] M. C. Anchahua, L. V. Garnique, and J. A. Tarazona, "User Experience Maturity Model for Ecommerce Websites," in *Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI)*, 2018, pp. 1–6.
- [34] R. Ali, O. Ibrahim, and M. Nilashi, "Loyalty of young female Arabic customers towards recommendation agents: A new model for B2C E-commerce," *Technol. Soc.*, vol. 61, no. May, p. 101253, 2020.
- [35] N. H. Chowdhury, M. T. P. Adam, and T. Teubner, "Time pressure in human cybersecurity behavior: Theoretical framework and countermeasures," *Comput. Secur.*, vol. 97, p. 101931, 2020.
- [36] A. Seffah, M. Donyae, R. B. Kline, and H. K. Padda, "Usability measurement and metrics: A consolidated model," *Softw. Qual J.*, vol. 14, no. 2, pp. 159–178, 2006.
- [37] G. Gronier, "Measuring the first impression: testing the validity of the 5 second test," *J. Usability Stud.*, vol. 12, no. 1, pp. 8–25, 2016.
- [38] N. Mohamad and N. L. Hashim, "UX Testing for Mobile Learning Applications of Deaf Children," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 294–299, 2021.
- [39] C. J. Van Staden, J. A. Van Biljon, and J. H. Kroeze, "Using a user experience evaluation framework for eModeration," in *2017 Conference on Information Communication Technology and Society, ICTAS 2017 - Proceedings*, 2017, pp. 1–6.
- [40] Z. Rasyida, M. C. Lam, A. B. Khairul Azmi, and I. Ahmad Khaldun, "User Experience Model for Remote Envenomation Consultation Mobile Application with Decision Support Ability," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 4-2, pp. 1470–1479, 2018.
- [41] M. S. A. B. A. Ghani and S. N. B. Wan Shamsuddin, "A Systematic Literature Review: User experience (UX) Elements in Digital Application for Virtual Museum," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 2801–2807, 2020.
- [42] A. I. Nur, H. B. Santoso, and P. O. H. Putra, "The Method and Metric of User Experience Evaluation: A Systematic Literature Review," in *International Conference on Software and Computer Applications*, 2021, pp. 307–317.

# Applying Artificial Intelligence in Retrieving Design Solution

Y. Moubachir, B. Hamri, S. Taibi

Q.S.M. Lab, Mohammed V University in Rabat, EMI, Morocco

**Abstract**—Design is a very important step in the product life cycle, because it is generally the key for the success or the failure of the product. The field of design theories and methodologies is full with theories and methods that have been taught and developed throughout the years. Most of them rely on subdividing the design process into phases, where the transition between each two phases rely on using some design tools. One of the main challenges of nowadays is to find a way for the integration of artificial intelligence (AI) in the design process. This integration could be very benefic, due to the fact that AI can learn quickly the relationship between input and output of any phenomena, and it can also give us a prediction of the behavior, if the inputs parameters vary. In our previous work we shaded the light, on how we can improve the transition between design phase by storing and retrieving design solutions using morphological analysis and design tools like DFX. In this work, we present a deferent methodology to perform this transition which rely on using an artificial intelligence tool called Artificial Neural Networks (ANN) instead of morphological analysis to retrieve the right design solution. To illustrate this method, we will take the same example from our previous work and will show how we can use ANN to learn and predict the right design solution.

**Keywords**—Artificial intelligence; ANN; design methodologies; DFX; morphological analysis

## I. INTRODUCTION

The design phase is a very important step in the product life cycle [1], even if it is generally costing for approximately 5% of the global cost of the project. But the decisions made in this step influence 70% of the global cost of the project [2]. Nowadays, product design is becoming a very challenging task due to the fact that a design team need to have a deep understanding of a variety of fields of knowledge (technological, social, cultural...) to ensure the success of their products once it's putted in the market [3]. In this scope the field of design theory and methodologies is rich of research and result that have been used and taught in industry and education [4], which are used as guidelines to help identifying right steps to take, for the purpose of identify the shape of the product that will succeed. Those facts lead to the growing complexity of products [5] which make the design process a very difficult step. This complexity is related to the fact that the number of parameters that we need to take in consideration in our product are increasing exponentially.

The application of Artificial neural network (ANN) in the design phase of the product life cycle is still in the earlier stage of development [6] [7]. And it's integration in the design process is systematically growing in many field of design [8]. This is due to the fact that ANN has a big potential to help

reducing complexity in the design process, and leading the designers to converge quickly to the wright design solution [9].

The goal of this work is to present a new methodology that will help in the integration of artificial neural network (ANN) in the design process. To achieve this goal, we will start from our previous work which had as purpose, to store and retrieve design solution based on morphological analyses and design tools [10], in that work the idea was to improve the transition between design phases, by relaying on storing technical solution based on some criterions that are already predefined, and then retrieving the right one based on the selected criterions, using morphological analysis method. In this work, the main contribution relay on using artificial neural network for retrieving design solution instead of morphological analysis. This can be achieved based on a deferent modelization of the problem. To illustrate this new method, two type of modelization of the problem will be proposed and compered.

In this scope our work is built up as bellows. In Section II, a short presentation of the main idea for the previous work, and the illustrative example that will be used. Section III, depicts the functioning and the sizing of the artificial neural network are presented. Section IV, is devoted to show the two types of possible modelization of the ANN that can be used to retrieve design solution works. Section V sums up our contributions and outlines some possible upcoming work.

## II. PREVIOUS WORK

The main idea of the previous work was to show how we can identify a suitable design solution based on the desired values of set of inputs criterions (Fig .1).

The illustrative example was the “mounting of spur gear in a shaft”, as a design problem where we have different possible solutions regarding the inputs criteria. For that example, eight design solutions were proposed as shown in Fig. 2 and each design solution has different characteristics:

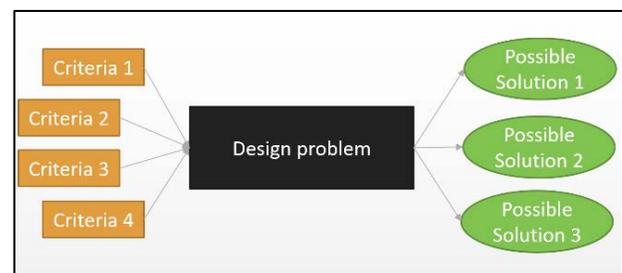


Fig. 1. Configuration of the Problem.

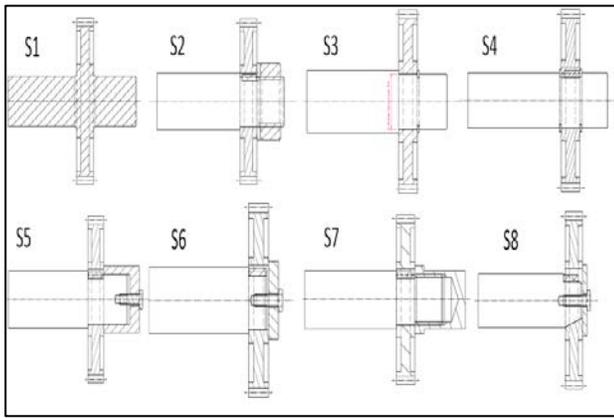


Fig. 2. Design Solutions for Mounting Spur Gear in a Shaft.

- S1: the spur gear is directly mounted in the shaft.
- S2, S4, S5, S6 and S7: A key is mounted between the shaft and the gear. The shaft and the spur gear hole have a cylindrical shape, with different axial blocking solutions.
- S3: An interference fit is used between the gear inner hole and the shaft.
- S8: A key is mounted between the shaft and the gear, and the shaft and the spur gear hole have a tapered shape.

For this illustrative example the developed criteria are Manufacturability index [11] (MI), Number of commercial part, Total number of part, manual assembly time [12] and Reparation cost index. For each solution we calculate the possible value regarding each preselected criterion. The results of calculations are given in Table I.

TABLE I. RESULT FOR EACH SOLUTION

| Solutions | MI   | Commercial parts | N° of parts | $(\alpha+\beta)/720$ | Reparation index |
|-----------|------|------------------|-------------|----------------------|------------------|
| S1        | 0.31 | 0                | 1           | 0                    | 1.00             |
| S2        | 0.30 | 3                | 5           | 0.5                  | 0.36             |
| S3        | 0.49 | 1                | 3           | 0.25                 | 0.41             |
| S4        | 0.55 | 3                | 5           | 0.5                  | 0.34             |
| S5        | 0.28 | 2                | 5           | 0.5                  | 0.35             |
| S6        | 0.32 | 2                | 5           | 0.5                  | 0.34             |
| S7        | 0.30 | 1                | 4           | 0.5                  | 0.41             |
| S8        | 0.14 | 3                | 5           | 1                    | 0.48             |

### III. ARTIFICIAL NEURAL NETWORK

The artificial neural network is a mathematical tool that was developed based on the functioning model of the human brain. This mathematical tool is used in many fields of artificial intelligence, like image recognition, machine learning, prediction and classification. In this work ANN will be used as machine learning tool, where the aim is to learn the obtained values of each solution, in order to build a mathematical model

which will give us for each set of criteria, the most suitable possible solutions.

#### A. Functioning of the Neural Network

An artificial neural network is a model that generally consists of three types of layers. An input layer, several hidden layers and an output layer (Fig. 3). For each neuron in the network (Fig. 4), the output value is calculated as follow (1):

$$x_j = \sigma\left(\sum_{i=1}^n w_{ij} \cdot O_i\right) \quad (1)$$

Where :

$x_j$  : The neuron output value.

$w_{ij}$  : Value of the weight which links two neurons.

$O_i$ : Neuron output value of the preceding layer with ( $O_n=1$  bias).

$\sigma$  : Transfer function.

There are several type of transfer function that can be used in the neural network, like the Gaussian function (2), tangent hyperbolic (3), linear (4) or the sigmoid function (5). This last one is generally the mostly used transfer function [13].

$$\sigma(t) = \exp(-t^2) \quad (2)$$

$$\sigma(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}} \quad (3)$$

$$\sigma(t) = t \quad (4)$$

$$\sigma(t) = \frac{1}{1 - e^{-t}} \quad (5)$$

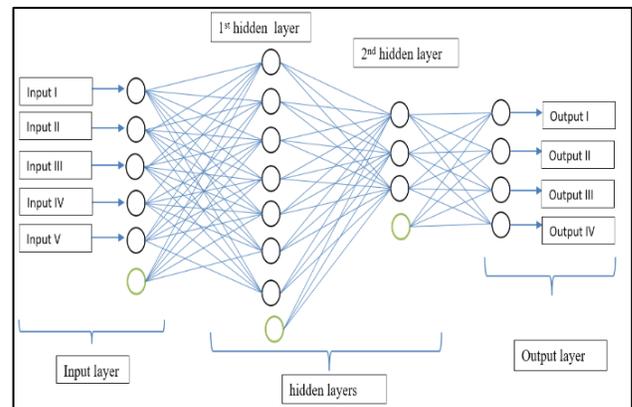


Fig. 3. Neural Network Model.

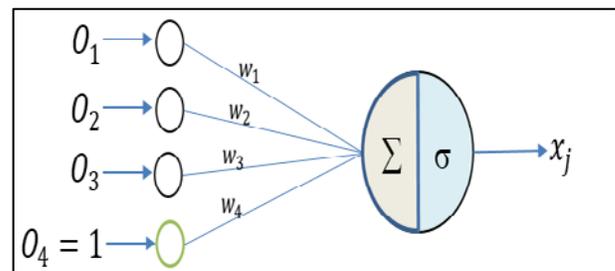


Fig. 4. The Model of the Activation of each Neuron.

### B. Sizing the Neural Network

The purpose of sizing the neural network is to identify the most suitable neural network for our case of study. This operation begins by identifying the number of neurons in each layer and the number of hidden layers. In other words, the goal is to find the smallest size of the network that has an equation that generalizes the best the relationship between inputs and outputs (Fig.5 (a)) to do so two things must be avoided:

- The overfitting: which means that the network used is too large, so that the equation obtained has integrated the errors in the training set of the model (Fig. 5(b)).
- The underfitting: which means that the network used is too small, so that the equation obtained is not adequate with the actual model (Fig. 5(c)).

Thus, to properly size the network, we will start with the smallest possible network. We will then calculate the error obtained. If this error is too large, we will continue to increase the size of the network until the error becomes too small.

### C. The used Neural Network Model

The use of a neural network begins with the learning process. The learning process begins by a set of input and output data are given to the network for learning purpose. In our case we will give the network, for each set of output values  $S_k$  (solution S) a set of input values corresponding to it  $I_m$  (criterion) where « m » is the number of inputs and « k » is the number output. The goal of this model is to identify from the training data the values of the  $w_{ij}$  of the mathematical equation which links the input values to output values. The learning of the network is done according to the following steps :

- 1) The identification of the input values and the corresponding output values (training set).
- 2) Initialization of the values of  $w_{ij}$ .
- 3) Calculating the error E : which is the square of the difference between the results of the neuron values and real the output (6).
- 4) Using the Gradient Descent Method (Several variants of the gradient decent are possible, in our example we will use the delta-bar-delta method [5] ) to modify the values of  $w_{ij}$ .
- 5) Return to step 3 until the error becomes small.

$$E = 1/2(\sum_i^k (S_i - x_i)^2) \quad (6)$$

With :

k : is the number of outputs.

$S_i$  : the value of the output of the training set.

$x_i$  : the value of the output obtained by the neural network.

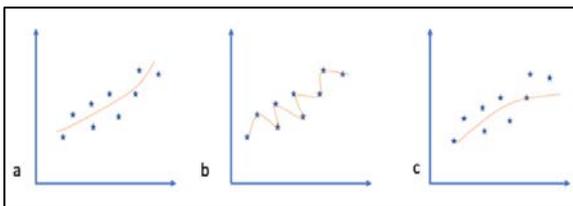


Fig. 5. (a) Proper Result, (b) Overfitting, (c) Underfitting.

## IV. LEARNING MODEL

In this section we present the two learning models for the studied problem “mounting a pinion in a shaft”. The first model consist of using a network of five inputs and three outputs, and the second one consist of using a network of five inputs and eight outputs. Then we will compare the obtained results of the two models.

In our case we will refer to the output values of the network by  $S_k$ , which refer to one of the eight design solutions presented in Fig. 2. For the set of input values related to the  $S_k$  solution, we will refer to them by  $I_k$ , this set is composed of 5 values which are corresponding to the values obtained by each of the  $S_k$  solution in regard to the five criterions  $I_k$ . For example,  $I_1$  related to the  $S_1$  solution is {0.3 ;1 ;0 ;1 ;0 ;1 }.

### A. First Model

For this first learning model, a neural network composed of 5 inputs and 3 outputs will be used. In this case the five inputs are corresponding to the 5 values of the criterions. The input values for each criterion should be standardized (there values should be between -1 and 1) [5], this step is very important so the network can give us reliable results. For this, we will divide the values of the number of commercial parts and the total number of parts by 10, and for the other criterions they are already standardized.

The output values are three neurons so each solution will be encoded as follow:

S1 corresponds to the value (0,0,0).

S2 = (0.0,1)

S3 = (0.1,0)

S4 = (0.1,1)

S5 = (1.0,0)

S6 = (1.0,1)

S7 = (1.1,0)

S8 = (1.1,1)

The python programming language and Tcl as graphical user interface (Fig. 6) were used to do the learning with a developed backpropagation algorithm. The weights values of our network were calculated using the Delta-Bar-Delta method [5].

To size the network, two possible networks were evaluated. the first network “5.2.3” consists of two neuron in the hidden layer and the second one “5.3.3” is composed of three neuron in the Hidden layer ( Fig. 7).

The details of the obtained results are for each network are:

1) For neural network “5.2.3”:

- Learning parameter: Learning rate: 0.1; Momentum: 0.7; Delta-bar-delta parameter: ;  $\kappa=0.05$ ;  $\theta=0.3$ ;  $\phi=0.2$
- Global error = 0.073.
- Weights: [array([[ -6.9558979, -10.7009645, 17.15070722, 9.81960633, -12.31982889, -

1.65349976],[ 1.7778629, -8.79354065, 17.27106976, -3.2645242 , 16.7249306 , -13.02356945]], array([[ 35.5968495 , 34.0311664 , -37.00298035],[ -5.08638478, -8.1145575 , 5.80673535],[5.01774992, -49.84129666, 4.49115069]]).

• Results

- [S1] : [ 4.48e-02 9.21e-02 2.28e-20] ≈ [ 0,0,0]
- [S2] : [ 1.69e-01 5.97e-01 9.71e-01] ≈ [ 0.0,1]
- [S3] : [ 2.67e-14 9.89e-01 8.41e-02] ≈ [ 0.1,0]
- [S4] : [ 5.08e-05 8.23e-01 8.54e-01] ≈ [ 0.1,1]
- [S5] : [ 9.83e-01 3.33e-01 4.27e-01] ≈ [ 1.0,0]
- [S6] : [ 9.67e-01 3.66e-01 6.00e-01] ≈ [ 1.0,1]
- [S7] : [ 8.39e-01 3.89e-01 8.63e-02] ≈ [ 1.1,0]
- [S8] : [ 9.27e-01 4.43e-01 9.74e-01] ≈ [ 1.1,1]

2) For Neural network“5.3.3”:

- Learning parameter: Learning rate: 0.1 ; Momentum: 0.7 ; Delta-bar-delta parameter:  $\kappa=0.05$ ;  $\theta=0.3$ ;  $\phi=0.2$ .
- Global error = 5.96 e-10.
- Weights :[array([[ 21.28907954, 7.92106131, -16.21637106, 9.40439058, 0.42501641, -5.98126962],[ -23.6018077, -14.94841144, 9.47059949, -11.06965724, 29.07147555, 9.45002567],[ -7.3583812 , 22.28383593, -1.5479036 , -16.2858285, 6.4895823, 3.67820648]], array([[ -22.88075349, 4.32776792, -33.44627925, 25.59618602], [ 37.39235073, 15.50577841, -22.68797391, -17.65045261],[ 36.05731457, -54.60337139, -8.63428378, 11.54696872]])].

• Results:

- [S1] : [ 8.05 e-06 1.13 e-05 1.51 e-17]
- [S2] : [ 3.18 e-05 1.87 e-09 9.99 e-01]
- [S3] : [ 4.68 e-06 9.99 e-01 2.34 e-05]
- [S4] : [ 2.55 e-05 9.99 e-01 1.00 e+00]
- [S5] : [ 9.99 e-01 1.81 e-06 4.69 e-05]
- [S6] : [ 9.99 e-01 4.71 e-05 9.99 e-01]
- [S7] : [ 9.99 e-01 9.99 e-01 8.74 e-13]
- [S8] : [ 9.99 e-01 9.99 e-01 9.99 e-01]

From this study it is noted that the network "5.3.3" provides a global error around 5.96 E-10 (Fig. 8) while the network "5. 2 .3" gives a global error (Fig. 9) of 0.073. So we can say that the network " 5.3.3 " is the most suitable for our example, because it is the smallest network that provides good results.

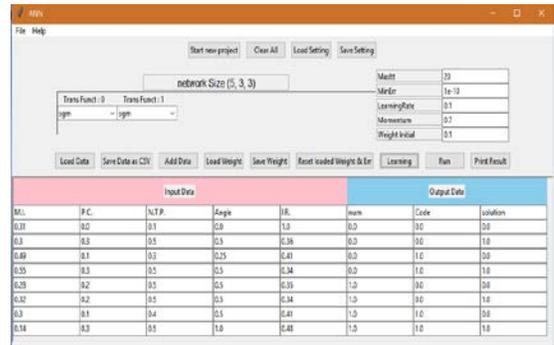


Fig. 6. The Developed Program for the Learning.

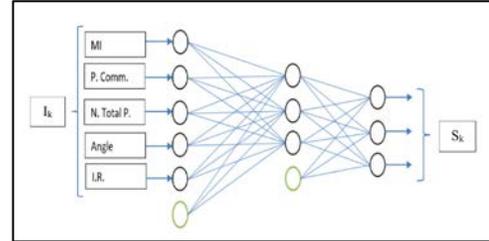


Fig. 7. The (5.3.3) Network used for the Learning of the Technical Solution.

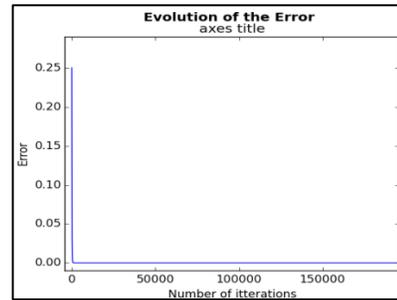


Fig. 8. The Evolution of the Error after each Iteration using a “5.3.3” Network.

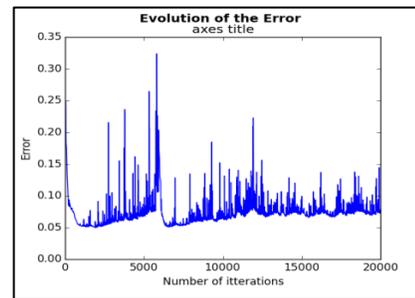


Fig. 9. The Evolution of the Error after each Iteration using a “5.2.3” Network.

B. Second Model

This second learning model is a neural network with 5 inputs and 8 outputs. The five inputs correspond to the values of the chosen criteria and the eight neurons were chosen for the output values. In this case, each output neuron corresponds to a given solution, if an output neuron obtains the value of 1, then this solution is the most suitable in regards to the given values of the inputs criteria.

To size this network, two networks were evaluated the first one is with one neuron in the hidden layer, and the second one with two neurons in the hidden layer (Fig. 10). Details of the obtained results are:

1) For Neural network“5.1.8”:

- Learning parameter: Learning rate: 0.1 ; Momentum: 0.7 ; Delta-bar-delta parameter:  $\kappa=0.05$ ;  $\theta=0.3$ ;  $\phi=0.2$ .
- Global error: 0.06801684134122636.
- Weights:[array([[ 3.66954535e-02, -1.02956629e+02, 7.11856022e+01, -2.84464136e+01, 3.84737574e-01, 2.04345583e+00]]), array([[ 3.05821270e+03, -3.05509115e+03], [ -5.17925747e+00, -6.91778960e-01], [ 5.05589009e+01, -5.12220784e+01], [ -5.17697254e+00, -6.91783745e-01], [ 1.76212801e+00, -3.19181346e+00], [ 1.76002122e+00, -3.18999900e+00], [ 5.05586156e+01, -5.12217948e+01], [ -1.55159832e+04, 4.93991111e+00]])].

• Results;

- [S1] : [ 9.48e-01 2.81e-03 3.39e-01 2.81e-03 1.93e-01 1.93e-01 3.39e-01 0.0]
- [S2] : [ 0.00 3.32e-01 5.87e-23 3.32e-01 3.95e-02 3.95e-02 5.87e-23 5.05e-03]
- [S3] : [ 3.64e-02 2.84e-03 3.16e-01 2.84e-03 1.92e-01 1.92e-01 3.16e-01 0.00]
- [S4] : [ 0.00 3.32e-01 5.87e-23 3.32e-01 3.95e-02 3.95e-02 5.87e-23 4.98e-03]
- [S5] : [ 2.16e-64 3.62e-03 4.15e-02 3.62e-03 1.80e-01 1.80e-01 4.15e-02 0.00]
- [S6] : [ 1.54e-64 3.62e-03 4.13e-02 3.63e-03 1.82e-01 1.80e-01 4.13e-02 0.00]
- [S7] : [ 3.64e-02 2.84e-03 3.16e-01 2.84e-03 1.92e-01 1.92e-01 3.16e-01 0.00]
- [S8] : [ 0.00 3.33e-01 5.68e-23 3.33e-01 3.94e-02 3.95e-02 5.68e-23 9.92e-01]

2) For Neural network“5.2.8”:

- Learning parameter: Learning rate: 0.1 ; Momentum: 0.7 ; Delta-bar-delta parameter:  $\kappa=0.05$ ;  $\theta=0.3$ ;  $\phi=0.2$
- Global error: 2.16e-08
- weights: [array([[-11.72500802, -13.56629037, 12.12204135, 20.69551574, 7.94683641, -11.8472096 ],[-22.54860196, -49.77196842, 0.89553372, 5.18327433, 17.64766934, 5.37958571]]), array([[ -23.52660555, 23.42517761, -13.33842946], [ 40.44144954, -435.47791596, -10.05392598], [-525.18616036, -19.43990113, 11.30272584], [ -43.84693704, -145.95717923, 10.66856872], [ 244.01849652, -103.39034649, -169.71850489], [ 153.29525942, -326.16577446, -78.25786574], [

23.84288311, 116.50855895, -125.77287833], [ 106.3662878 , -6.67651725, -92.49732366]])]

• Results:

- [S1]: [ 9.99e-1 3.46e-194 1.14e-004 1.62e-59 3.80e-119 3.02e-176 9.88e-5 1.02e-43]
- [S2] : [ 3.37e-11 9.99e-1 1.74e-100 6.81e-5 7.50e-26 2.64e-4 1.50e-50 1.05e-19]
- [S3] : [ 3.83e-5 7.49e-31 9.99e-1 9.42e-5 1.91e-80 5.95e-54 1.92e-48 3.02e-41]
- [S4] : [ 6.72e-7 1.92e-4 2.73e-4 9.99e-1 1.69e-70 3.05e-32 5.78e-55 3.50e-39]
- [S5] : [ 5.09e-13 4.66e-21 3.69e-178 3.15e-21 9.99e-1 2.81e-4 3.93e-39 1.07e-4]
- [S6] : [5.91e-13 2.45e-4 1.30e-153 4.96e-13 4.74e-4 9.99e-1 3.60e-45 3.23e-9]
- [S7] : [ 3.017e-5 3.01e-175 6.42e-198 2.61e-74 2.59e-28 9.33e-117 9.99e-1 1.77e-4]
- [S8] : [1.37e-8 4.90e-139 1.14e-230 6.47e-66 1.98e-4 1.29e-81 1.84e-4 9.99e-1]

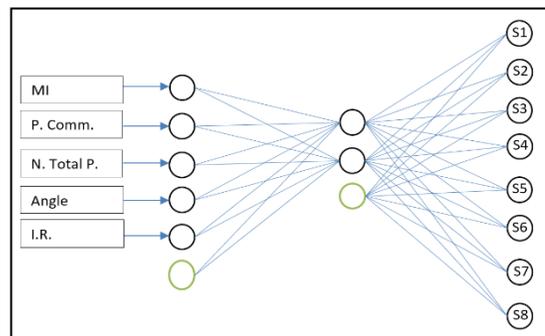


Fig. 10. A (5.2.8) Net used for the Learning of the Technical Solution.

According to this study, we notice that the network “5.2.8” gives a total error of about 2.16E-08 (Fig. 11) while the network “5. 1. 8” (Fig. 12) gives a total error of 0.068. So we can say that the network “5. 2. 8” is the most suitable for our example.

C. Synthesis

As a summary, we can see that the first network “5.3.3” is a network that was able to record the relationship between the elements of input and outputs, with a network that contains fewer parameters (13 neurons). But the downside of this model is that we are unable to identify the nearest solution if we deviate the values of the inputs criteria. On the other hand, the network “5.2.8” is a network which contains more parameters. But in the other hand, it has the advantage of giving the nearest solution if we deviate the values of the inputs criteria.

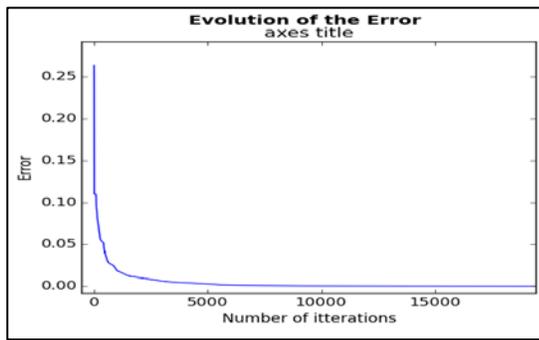


Fig. 11. The Evolution of the Error after each Iteration using a “5.2.8” Network.

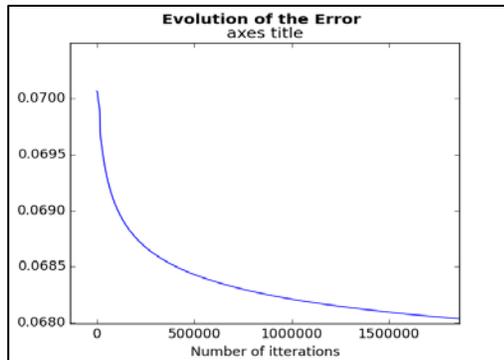


Fig. 12. The Evolution of the Error after each Iteration using a “5.1.8” Network.

## V. DISCUSSION AND CONCLUSION

The use of neural network method in the retrieval of the wright design solution has a great practical value to industry, because it has the potential of saving time and cost in the design processes [14].

In comparison to our previous work [6], where we used morphological analysis to retrieve design solution, we can see that the use of ANN has the advantage of predicting possible solutions if we change the values of the inputs criteria. Where in our previous work, the values of the criteria were fixed and cannot be changed. In this case, if our design problem had different values for at least one of the criteria the method cannot be used, but in the other hand the use of ANN help us to overcome this issue.

The purpose of this work is to present a method to improve the identification of technical solutions that will meet a set of design parameters that were pre-identified in the previous phase of the design. The proposed method is an idea that can be applied in several other problems like, selection of bearing type, Material selection or the choice of the machining process. The advantage of this method is that, it allows us to make a quick and visual choice of the suitable solution by using the power of ANN. But the drawback is that this method requires a huge work in advance to include all kinds of possible solutions in regards to the developed criteria.

## REFERENCES

- [1] K. Ulrich, DESIGN: Creation of Artifacts in Society. the University of Pennsylvania., 2011.
- [2] S. Munro, D. Foreman, D. McCarthy, I. Chambers, and Thomas, “Lean Design : Value Quality Profit,” Munro & Associates, Inc., 2013.
- [3] D. G. Ullman, The Mechanical Design Process, Fourth Edition. McGraw-Hill Higher Education, 2010.
- [4] T. Tomiyama, P. Gu, Y. Jin, D. Lutters, C. Kind, and F. Kimura, “Design methodologies: Industrial and educational applications,” CIRP Ann. - Manuf. Technol., vol. 58, no. 2, pp. 543–565, Jan. 2009.
- [5] W. ElMaraghy, H. ElMaraghy, T. Tomiyama, and L. Monostori, “Complexity in engineering design and manufacturing,” CIRP Ann. - Manuf. Technol., vol. 61, no. 2, pp. 793–814, Jan. 2012.
- [6] H. A. Khayyat, “ANN based Intelligent Mechanical Engineering Design: A Review,” Indian J. Sci. Technol., vol. 11, no. 27, pp. 1–7, 2018.
- [7] Z. Zhang and K. Friedrich, “Artificial neural networks applied to polymer composites: A review,” Compos. Sci. Technol., vol. 63, no. 14, pp. 2029–2044, 2003.
- [8] W. Sitek and J. Trzaska, “Practical aspects of the design and use of the artificial neural networks in materials engineering,” Metals (Basel), vol. 11, no. 11, 2021.
- [9] I. M. L. Ferreira and P. J. S. Gil, “Decision Support Tool for Conceptual Design using Neural Networks,” 2010, no. 1.
- [10] Y. Moubachir and D. Bouami, “Storing and Retrieving Design Solution in the Physical Domain Based on DFX Tools and Morphological Analysis,” Procedia CIRP, vol. 34, pp. 64–68, 2015.
- [11] S. K. Ong, M. J. Sun, and a. Y. C. Nee, “A fuzzy set AHP-based DFM tool for rotational parts,” J. Mater. Process. Technol., vol. 138, no. 1–3, pp. 223–230, Jul. 2003.
- [12] Z. Yoosufani, M. Ruddy, and G. Boothroid, “Effect of part symmetry on manual assembly times,” J. Manuf. Syst., vol. 2, no. 2, pp. 189–195, 1983.
- [13] S. Samarasinghe, “Neural Networks for Applied Sciences and Engineering,” p. 581, 2007.
- [14] I. Article, “NIRS: Large Scale ART-I Neural Architectures for Engineering Design Retrieval,” neural Des. Retr., vol. 7, no. 9, pp. 1339–1350, 1994.

# State-of-the-Art Approach to e-Learning with Cutting Edge NLP Transformers: Implementing Text Summarization, Question and Distractor Generation, Question Answering

Spandan Patil<sup>1</sup>, Lokshana Chavan<sup>2</sup>, Janhvi Mukane<sup>3</sup>, Dr. Deepali Vora<sup>4</sup>, Prof. Vidya Chitre<sup>5</sup>  
Department of Information Technology, Vidyalankar Institute of Technology, Mumbai, India<sup>1,2,3,5</sup>  
Department of Information Technology, Symbiosis Institute of Technology, Pune, India<sup>4</sup>

**Abstract**—Amid the worldwide wave of pandemic lockdowns, there has been a remarkable growth in E-learning. Online learning has become a challenge for students. It has become difficult for students to find the content they need. The mounting accessibility of textual content has necessitated comprehensive study in the areas of automatic text summarization and question generation. Multiple Choice Questions is very smooth for evaluations, and its assessment is implemented through computerized applications in order that results may be declared within some hours, and the evaluation system is 100% pure. The system proposes an interactive reading platform where the user can upload an E-Book and get textual summary and generates questions like MCQs, fill in the blanks and one word. The user can also evaluate the questions answered. The proposed system is an all-in-one interactive reading platform.

**Keywords**—Machine intelligence; natural language processing; neural networks; predictive models; text processing

## I. INTRODUCTION

The “Live with Covid” era has notably modified the manner we live. The field of education can't isolate itself from the drastic adjustments, resulting in the near-total closures of schools, early childhood education and care (ECEC) services, colleges, and universities [1]. It has compelled us to follow the online mode of learning. As transferring to online learning has introduced us to flexibility and self-paced mastering. But there are few cons to this. The new format of classes for students has left them with a lack of motivation and creates a sense of isolation from the classrooms which may affect their academic performance throughout the term. Also, with the development of information technology, more and more information appears on the internet, retrieving the needed information and making sense out of huge data becomes difficult for users. Hence, there comes a need for a system which can provide us with summarization and question generation for easier and quicker retrieving of relevant information from huge chunks of data and to test the understanding of the subject through assessments.

In Section II, Review and Planning of the paper is discussed. In Section III, various text processing algorithms like LSTMs, T5, BERT, WordNet, ConceptNet, Sense2Vec, etc. are discussed and the best suited ones are elaborated. In

Section IV, Literature Survey was carried out wherein, technical research papers and some existing systems were studied. The gaps in the theory and applications are also addressed. In Section V, Inferences from the literature survey are mentioned. In Section VI, the Proposed System is described in detail along with its workflow and features. In Section VII, the implementation part of the system is discussed in detail. In Section VIII, the results are presented along with various comparisons between the findings. In Section IX, conclusions are stated and possible topics for future research are mentioned.

## II. REVIEW AND PLANNING

Text Processing is one of the most common tasks in many ML applications. The review considered following queries.

Q1- What are the different techniques for performing the NLP tasks like Text Summarization, Question Generation and Question Answering?

Q2- What are the different distractor generator algorithms used to generate three distractor options in MCQs?

Q3- What challenges were faced while using these algorithms? And which one was the best suited for the use case?

For the survey, databases of IEEEEXPlore, Google Scholar, and Articles were searched manually, by using various keywords like “Text Summarization”, “Question Generation”, “Question Answering”, “Distractor Generation”, etc. The search was narrowed down to research that only perform Abstractive Text Processing [2] which includes the tasks like text summarization, Question generation and question answering. The approaches to generate distractors for incorrect options of MCQs were also studied. Research papers of various Text Processing Approaches were studied from the mentioned repositories like Journals, Conferences and Articles. While trying to select the literature for the system, time duration was limited from 1997 to 2021. Along with Research Papers certain existing system were also reviewed. The research papers that satisfy the above conditions were studied and a few comparisons were made based on certain parameters. The notable points are highlighted in this paper. These remarks helped in identifying solutions to the review questions.

### III. TEXT PROCESSING ALGORITHMS

Text Processing is one of the most common tasks in many NLP applications [3]. These algorithms help the computers to analyze, understand and derive meaning from human language in a smart and useful way. For this system, out of all the Text Processing Algorithms, RNN based Sequence model like LSTM and Transformer based models like T5 and BERT were studied. The highlights from research papers are mentioned below.

#### A. LSTM versus Transformers

LSTM were one of the most popular choices for performing Natural Language processing tasks [4], but were replaced after the introduction of current state of the art transformer which surpass LSTM over the accuracy and convenience of performing the NLP tasks [5]. Limitations of LSTM are- it is difficult to train (takes very long time), transfer learning never really worked, and it must be computed in serial per token [6].

#### B. T5 Transformer

The text-to-text framework introduced in the paper [7], allows NLP tasks, like document summarization, machine translation, question answering regression tasks to be trained to predict the string representation of a number instead of the number itself using the same model for loss function, and hyperparameters. The input and output of the T5 model is always purely text to text format i.e., text string as shown in Fig. 1.

#### C. BERT Transformer

The BERT model in [8], being inspired by the Cloze task (Taylor, 1953) alleviates the unidirectionality constraint by using a “Masked Language Model” (MLM) pre-training objective. Few of the input tokens are arbitrarily masked by the MLM and their original vocabulary id is predicted solely based on the context. It also makes use of “next sentence prediction” task in addition to mask language model to jointly pretrain text-pair representations as shown in Fig. 2.

#### D. BART Model

BART is a denoising autoencoder used for pre-training sequence-to-sequence models. It is trained by corrupting text with random noise function thus, model learns to restructure the original text as shown in Fig. 3. A standard Transformer with simple neural machine translation architecture is used in BART. It evaluates several noising approaches. It finds the optimum performance by arbitrarily shuffling the sequence of original sentences and thereby uses a novel in-filling scheme, where a single mask token is placed in spans of text.

#### E. Distractor Generator for Incorrect Options

For distractor generation mainly WordNet, ConceptNet, Sense2Vec were studied.

WordNet is a large lexical knowledgebase of English. Every word (i.e., Adjectives, nouns, verbs) is grouped into sets of logical synonyms (synsets), expressing a unique concept. Every Synset is interlinked to another by lexical relations and conceptual-semantic [10]. A hypernym is a higher-level category for a given word. Considering an example as shown

in Fig. 4, color is hypernym for red. Hyponyms are the sub-categories of an entity. A hyponym is a type-of relationship with its hypernyms [11]. A Co-Hyponym are words that shares the same hypernym as another word. To generate distractors the main goal is to extract co-hyponyms [12].

ConceptNet is a semantic network that is used to help computers understand the meaning of words that people use. It generates distractors for locations, items, etc. which have a “Part of” relationship [13]. ConceptNet Number batch is a set of semantic vectors, also known as word embeddings which can be used as direct representation of word definitions or an initial state for further machine learning [14]. Fig. 5 depicts an example of Generation of distractors using ConceptNet.

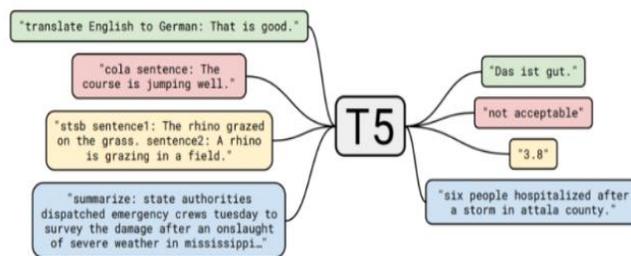


Fig. 1. T5 Text-to-Text Framework. [7].

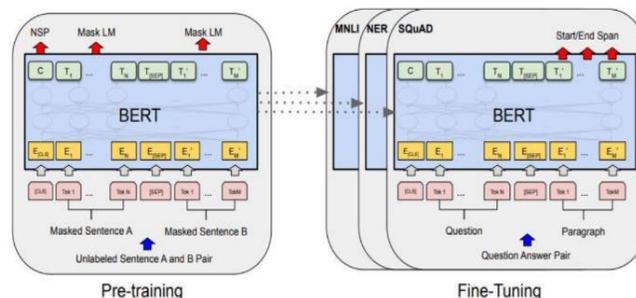


Fig. 2. BERT Fine-Tuning Procedure. [8].

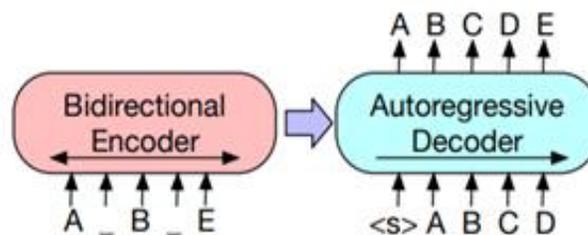


Fig. 3. A Schematic Diagram of BART. [9].

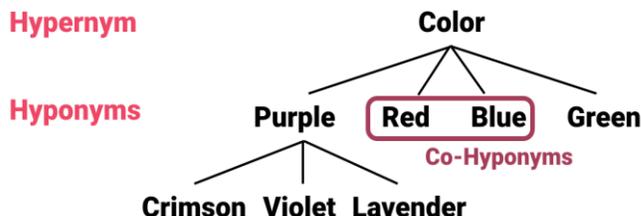


Fig. 4. An Example of Relationship between Hyponyms and Hypernyms.

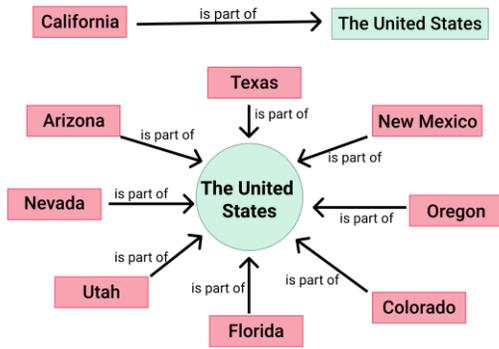


Fig. 5. An Example of Generating Distractors using ConceptNet.

Sense2Vec automatically generates relations among words from a text corpus in contrast to being human-curated. To predict a focus word given other words or to predict surrounding words of a given focus word a neural network algorithm is trained with millions of sentences as its dataset. Thus, resulting word vectors which are fixed size vectors or array representation of every word. The associations between different kind of words are represented by these word vectors, thus preserving the relationship among various words. The system uses 2015 Reddit vectors instead of the 2019 as the output obtained was slightly better.

#### IV. LITERATURE SURVEY

For the system, literature survey was conducted into two parts. For the first part of literature survey, several research papers related to text summarization, question generation and question answering were studied. The inferences from paper reading are tabulated in Table III. Later, study of existing systems was carried out that aid in the process of interactive reading experience and self-assessment. The observations from existing systems' review are tabulated in Table IV.

#### V. SURVEY INFERENCE

In the mentioned reviews, out of all the Text Processing Algorithms, RNN based Sequence model like LSTM and Transformer based models like T5 and BERT were studied. T5 is an integrated text-to-text model with text strings as its input and output, whereas BERT-style models generate outputs as a class label or a span of the input.

The issue of understanding each word based on the understanding of previous words couldn't be handled by traditional neural networks. Thus, Recurrent Neural Networks were introduced to handle the same. These networks have loops in them, allowing information to persist. The limitations of traditional RNN are that computation is slow because of the concurrent nature. If relu or tanh are used as activation functions, it becomes very difficult to process longer sequences. It is vulnerable to issues such as exploding and gradient vanishing. Further LSTMs came into picture. LSTMs are a special kind of RNN, capable of learning long-term dependencies and work well on a large variety of problems. Transformer became a huge achievement over the RNN based seq2seq models. Using transformer All to All comparison can be done fully parallel, it has multi-headed attention and

positional encoding and Transfer Learning worked well on it. But its limitations include attention can solely deal with text sequences whose size is pre-defined. The sequence must be split into fixed-sized segments or chunks before being given as input into the system. Fig. 7 depicts evolution of text processing algorithms.

Hence, due to its advantages in terms of speed and compatibility to the task, the different transformer models like T5, Distil BERT, Distil BART were decided to be used in the system for performing text summarization, question answering and question generation tasks as shown in Fig. 6.

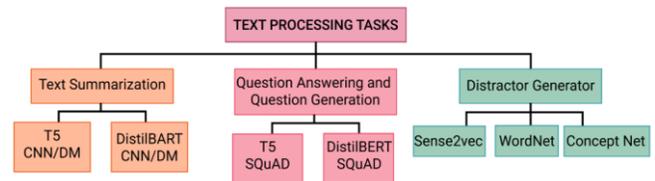


Fig. 6. Categories of Text Processing Tasks.

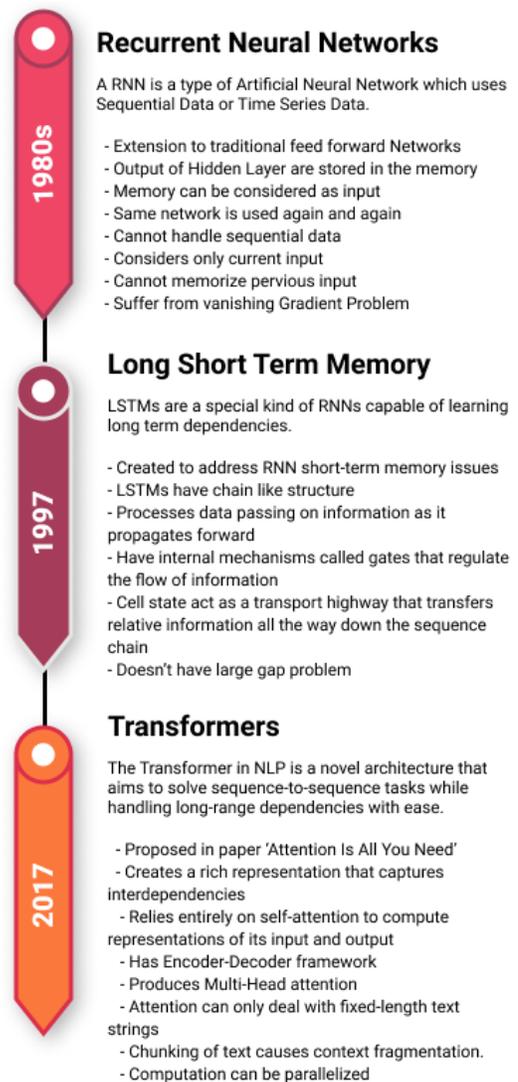


Fig. 7. Evolution of Text Processing Algorithms.

## VI. PROPOSED SYSTEM

Firstly, the user needs to sign up to the system and using credentials, log in to the home page. They can reset their passwords and edit their profiles as per their wish. The users will be categorized into guests and authors.

When a user (guest) authenticates into the system, an application tutorial will be displayed to make them familiar on how to use various features of the system. The user will either choose an E-Book (pdf) from the system library or upload one of their own. Then, the user will be able to read the E-Book in an E-reader, also have access to page wise summary and a self-assessment of the E-Book. The assessment will have three categories of questions – MCQ's, Fill in the blanks and One-word type questions. Additionally, the user can generate summarization from series of pages and get solution to a question they ask in context to a specific page. The system will display a collection of top-rated E-Books scraped from an E-Book rating and cataloguing website.

When authenticated authors logs into the system, they will be allowed to publish their E-Book (pdf), auto-generated page wise summary and a self-assessment of the E-Book to the library. The author will be able to search for their published E-Book from the library and get a preview of the generated summary and self-assessment. Furthermore, if they wish they can modify the same. Fig. 12 depicts proposed system.

## VII. IMPLEMENTATION

The systems' frontend designing is done using React.JS and in the backend, node.js express is used to host the web server and mongo DB Atlas is used for the database [15].

In the Mongo DB Atlas, a FLIP database is created in which there are three collections.

- User's E-Books: It stores all the data of the E-Book uploaded by the user.
- FLIP library: It stores all the data of the E-Books available in system's library.
- Authorized Authors: It stores the user's IDs of all the users that are categorized as authors.

MongoDB was chosen for its flexibility, scalability and cloud storage services. Also, it eases the restriction of a schema for of DB.

In the node server, APIs are available for uploading the E-Book to the server, performing CURD operations on the Mongo DB. It also runs three Python Scripts to perform mainly three tasks which are-

### A. Summary Generation and Self-Assesment Script

Firstly, for summary generation, T5 for Conditional Generation and T5Tokenizer from T5 base is used. This model was chosen over its other competitors like BERT, because it's pretrained on the much large and cleaner C4 dataset and in comparison, (base version), it contains nearly twice the number of parameters as BERT (T5: 220M & BERT: 110M) [7]. Additionally, it was also pre-trained specific for the text summarization task so no further fine tuning was required [8].

The content is then Pre-processed for removing all white spaces and is passed through the T5Tokenizer to get it tokenized. Its limitation is that there is a maximum limit of 509 tokens (excluding special tokens) for generation of summary. It was overcome by extracting the first chunk of 509 tokens from the tokenized content and then special tokens were added to them. Further, they are passed to the T5ForConditionalGeneration model to generate the summary whose length must be between 100 to 508 tokens. This generated summary is then added back to the front of the tokenized content. Now, the above process is repeated till tokenized content has less than 509 tokens. What this essentially does is retains the context of the previously generated summary with addition to the context provided by the additional tokens added in current iteration. Repeating this process will shorten down the size of the tokenized content to 509 tokens or less so that they can be fed to the T5 model at once without losing much context of the previous iteration.

An alternative approach to this, would be using the pre-trained DistilBART model. The process would majorly remain unchanged except for the limit for the length of the token chunk would be increased from 509 to 1022.

Next, Extraction of keywords is done using NER (Name Entity Recognition). Basic advantage of this method over other keyword extraction algorithms like Multipartite Rank, Tfidf, TextRank, etc is that it is able to identify Named Entities (NEs) which are real-life objects that are proper names and quantities of interest. And heuristically, when selecting answers for MCQ or other type of question in non-language related subjects these Named Entities have shown to provide more relevant and correct questions.

The name entities are extracted using SpaCy library from the content to get a list of keywords to be used as answers to the questions generated [16]. Using Maximal Marginal Relevance (MMR), the top five most relevant name entities can be procured out of the extracted ones. In MMR, the keywords that are most analogous to the text are selected. Then, iteratively new candidates are selected such that they both are analogous to the text and not analogous to the previously selected keywords. The similarities are measures based on Cosine similarity [17].

An alternative to MMR is Max Sum Similarity (MSS), The maximum sum distance of a pairs of data is calculated as the maximized distance which exists in between the two data points. In this case, candidate similarity was expected to be maximum to the document while minimizing the similarity between candidates. But a drawback in this is that, to get more diverse options there is a need to provide larger number of keywords to filter from, which is not possible for this system every time.

Then, for Generation of the question pre-trained T5ForConditionalGeneration base model is taken and it is fine-tuned on the question generation task using the SQuAD-The Stanford Question Answering Dataset [18]. The construction of fine-tuning dataset is done in the form of 3 columns the context, the answer and the question. The input is provided in the format of

```
context : (-context-) answer : (-answer-) </s>
question : (-question-) </s>
```

and provide the target in the following format-

In this way, nearly 80,000 rows are trained as a part of training dataset and 10,000 rows are used for validation of the trained model. For training the model batch size of 4 is used and for 1 epoch (which nearly takes about 4 hours to complete on the google collab notebook). Now, iteration over the keyword list is carried out and this fine-tuned model is used to generate questions for each keyword as output given the input as “context: (--summary--) answer: (--keyword--) </s>”.

To generate the distractors for the keywords the Sense2Vec Reddit 2015 is used. As Sense2Vec performs better with Named Entities compared to other algorithms like Wordnet, ConceptNet, etc.

The list of keywords is iterated and then- the best sense for selected keyword is generated and then the top thirty most similar words to it are found based on the best sense it gets. Then, the list of these words is filtered using Normalized Levenshtein Distance with a threshold of 0.7. Further, to select the top three distractors the MMR is used, by comparing the selected keywords and the distractors that are found. If any distractors for the keywords are not found, then those question-Keywords pair are used for one-word type questions.

Now, for generating fill in the blank’s questions, the top three keywords to be used as answers to the questions are found using the Multipartite Rank algorithm in the Python keyword extraction library as it seems to work best for these types of tasks [19]. Using the keyword processor in the FlashText library, the keywords are mapped with the sentences of which they’re a part of and then this Sentence-Keyword pair are used as Fill in the Blanks questions.

### B. Summary Generation for Collection of Pages Script

This script takes concatenated content of series of pages as input and then summarizes them using procedure same as that of the Summary Generation part mentioned in the Summary generation and self-assessment script.

### C. Question Answering for a Question asked in Context of a Specific Page Script

In this, pre-trained T5ForConditionalGeneration base model is taken and fine-tuned on the question answering task using the SQuAD- The Stanford Question Answering Dataset.[18] The fine-tuning dataset is constructed in the form of 3 columns- the context, the answer and the question. The input is provided in the format of “context: (--context--) question: (--question--) </s>” and provide the target in the following format of “answer: (--answer--) </s>” In this way nearly 80,000 rows are trained as the part of training dataset and 10,000 rows are used for validation of the trained model. For training the model batch size of 4 is used and for 1 epoch (which nearly takes about 4 hours to complete on the Google collab notebook). Now, this fine-tuned model is used to generate answers for the input questions, context which are fed to the model in the following format of “context: (--page\_content--) question (--input\_question--) </s>”.

An alternative approach to this would be using the pre-trained DistilBERT model. The process of fine tuning would majorly remain unchanged except for the input format would become “CLS (--question--) SEP (--context--) SEP” and for target is “CLS (--answer--) SEP”.

In the frontend, for sign up and sign in, Google Firebase Authentication Services are used. Thus, enabling the users to register using an email ID, password. And then signing in using the credentials provided to them. Further they are also provided with user support like resetting the password and updating the profile [20].

To upload an E-Book to the server, an upload API is used which in turn uses the node.js express- fileupload package and triggers the Summary generation and self-assessment Script where the summary and self-assessment for each page of the E-Book is generated and then the Inserting API of the CURD APIs is used to upload the generated content to the Mongo DB’s user’s E-Book collection. To retrieve the generated content of the E-Book uploaded by the user, the Retrieving API of the CURD APIs is used which uses E-Book name and user ID as a query in the Mongo DB user E-Book collection. To generate the summary for series of pages, the Summary Generation API is used which requires the content (concatenation of the content of series of pages) as request parameters. The API triggers Summary generation for Collection of pages Script and gives the script received content as an input. To generate answers to the questions in context to a specific page, the Question Answering API is used which requires the content (content of the specific page), question as request parameters. The API triggers Question Answering Script and provides the script with received content and question as input. For authors to upload an E-Book to the FLIP library, the library upload API is used which is same as upload API but the difference being that it uploads the E-Book to the FLIP library collection instead of the user’s E-Book collection. For authors to modify the content of the selected E-Book, the Modification API is used which uses the update API of the CURD APIs to update the data of the E-Book in the FLIP library collection.

## VIII. RESULTS

### A. Text Summarization

This task was performed using two approaches namely Pre-trained T5 model and Pre-trained DistilBART model. The performance of the two was evaluated based on the evaluation benchmarks like CNN/DM-ROUGE-1, CNN/DM-ROUGE-2 and CNN/DM-ROUGE-L which are recorded in the Table I.

1) *WMT 2016*: It is a group of datasets which can be used in the shared tasks - IT domain translation, an automatic post-editing, news translation, biomedical translation, etc. [21]. The score for WMT English to Romanian and Romanian to English are referred to as En-Ro and RO-EN [22].

2) *CNN/ daily mail*: It is a text summarization dataset. From CNN and Daily Mail Websites news stories, human generated abstractive summary bullets were generated as questions (with one of the entities hidden), and news stories as the corresponding passages which are used by the system to

answer the fill in the blank question as expected. The websites were crawled, using scripts released by authors and these scripts also extracted and generated pairs of passages and questions [23].

3) *ROUGE*: Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a collection of metrics used for evaluating natural language processing tasks like text summarization, machine translation software. An auto-produced summary or translation is compared against a reference or set of references (human-produced) summary or translation by the metrics.

- Rouge-N: Overlying of N-grams among the reference summaries and the system.
- Rouge-L: Longest Common Subsequence primarily totally based statistics. It takes into consideration sentence level structure similarity clearly and identifies longest co-occurring in collection N-grams automatically [24].

### B. Question Answering

This task was performed using two approaches namely Pre-trained T5 model and Pre-trained DistilBERT model. The performance of the two was evaluated based on the evaluation benchmarks like GLUE and SQuAD which are recorded in the Table II.

1) *GLUE*: The General Language Understanding Evaluation (GLUE) benchmark is a set of resources used to train, evaluate, and analyze Natural Language Understanding systems. It is of the model-agonistic format and hence, any system that is capable to process sentence and sentence pairs or outputting corresponding predictions is eligible to participate. Its' final goal is to drive analysis within the development of general and robust NLU systems [25].

2) *SQuAD*: The Question Answering Datasets use two major metrics called SQuAD Exact Match (EM) and SQuAD

F1 Scores. SQuAD EM is a simple yet strict all-or-nothing metric wherein every question-answer pair, if the characters of the model's prediction exactly match the characters of (one of) the True Answer(s), EM = 1, otherwise EM = 0. SQuAD F1 score is metric used for classification problems, and QA. It is ideally used when precision and recall are of equal importance It is calculated over individual words within the prediction against those in the True Answer. The premise of F1 score is that the number of shared words between the truth and prediction: precision is the ratio of the quantity of shared words to the overall number of words in the prediction, and recall is the ratio of the count of shared words to the total count of words in ground truth [26].

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

### C. Question Generation

For gaining knowledge about the accuracy of the system trials were carried out on 5, 10, 15 and 20 samples of E-Books each five pages long. The results of this approach on the system are recorded in the Table V and Table VI.

TABLE I. PERFORMANCE BASED ON EVALUATION BENCHMARKS FOR TEXT SUMMARIZATION

|                   | WMT EnRo / RO-EN | CNN/DM-ROUGE-1 | CNN/DM-ROUGE-2 | CNN/DM-ROUGE-L |
|-------------------|------------------|----------------|----------------|----------------|
| <b>T5</b>         | 28.0             | 42.05          | 20.34          | 39.40          |
| <b>DistilBERT</b> | 37.96            | 44.16          | 21.28          | 40.90          |

TABLE II. PERFORMANCE BASED ON EVALUATION BENCHMARKS FOR QUESTION ANSWERING

|                   | GLUE | SQuAD-EM | SQuAD-F1 |
|-------------------|------|----------|----------|
| <b>T5</b>         | 82.7 | 85.44    | 92.08    |
| <b>DistilBERT</b> | 77.0 | 77.7     | 85.8     |

TABLE III. STUDY OF RESEARCH PAPERS

| Paper Title                        | Key Points                                                                                                                                                                                                                                                                                                                                               | Drawbacks                                                                                                                                                                                        |
|------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Sepp Hochreiter, et al. [4]</b> | <ul style="list-style-type: none"><li>• LSTM being local in time and space having a complexity of O(1) per time weight and step.</li><li>• LSTM performs higher number of successful runs, learns faster</li><li>• It is efficient in solving complex, artificial tasks with long time lags</li></ul>                                                    | <ul style="list-style-type: none"><li>• It is difficult to train (takes very long time).</li><li>• Transfer learning never really worked, and it must be computed in serial per token.</li></ul> |
| <b>Colin Raffel, et al. [7]</b>    | <ul style="list-style-type: none"><li>• Explore transfer learning to introduce a unified text-to-text framework.</li><li>• Comparison between various aspects on masses of NLU tasks.</li><li>• Inference from C4 exploration and scale, advanced results on NLP tasks were recorded.</li></ul>                                                          | <ul style="list-style-type: none"><li>• Size of T5 model is 30 times more than the general NLP models</li><li>• It is expensive to use on commodity GPU hardware.</li></ul>                      |
| <b>Jacob Devlin, et al. [8]</b>    | <ul style="list-style-type: none"><li>• BERT a pre-trained deep bidirectional representation transformer model</li><li>• Performs pre-training on unlabeled text by jointly conditioning on right direction and left direction of context in all layers.</li><li>• Presents a MLM technique, for carrying out bidirectional training of models</li></ul> | <ul style="list-style-type: none"><li>• The fine-tuning and pre-training are inconsistent.</li><li>• The model file is too large, and the training time is too long.</li></ul>                   |
| <b>Mike Lewis, et al. [9]</b>      | <ul style="list-style-type: none"><li>• Proposes pre-training objective for sequence-to-sequence models as denoising autoencoder and uses Transformer architecture.</li><li>• Training done by corrupting textual content along with arbitrary noise function and the Language Model denoises it.</li></ul>                                              | <ul style="list-style-type: none"><li>• Outputs are highly abstractive with few copied phrases.</li><li>• Model has tendency to hallucinate unsupported information.</li></ul>                   |

TABLE IV. STUDY OF EXISTING SYSTEM

| Product                               | Key Points                                                                                                                                                                                                                                                                                                                                                                                                           | Drawbacks                                                                                                                                                        |
|---------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Text Summarization [27]</b>        | <ul style="list-style-type: none"> <li>Based on advanced NLP and ML technologies.</li> <li>Summarizes text from the URL or provided document</li> <li>Can be easily used in any environment</li> <li>Capable of making HTTP requests</li> </ul>                                                                                                                                                                      | <ul style="list-style-type: none"> <li>The system does not allow to upload whole E-Book.</li> <li>It does not have question generation functionality.</li> </ul> |
| <b>Automatic Text Summarizer [28]</b> | <ul style="list-style-type: none"> <li>A multilanguage Text Summarizing and Paraphrasing AI Tool</li> <li>Uses specific algorithm for extracting key points and uses extraction-based summarization</li> <li>Accessible by an API and is still in its development phase</li> </ul>                                                                                                                                   | <ul style="list-style-type: none"> <li>Upload of whole E-Book is not possible</li> <li>No option of question generation</li> </ul>                               |
| <b>Quillionz [29]</b>                 | <ul style="list-style-type: none"> <li>AI-powered platform for creating questions, quizzes and notes, allows to edit those questions and notes.</li> <li>Highlights important parts, summarized points, reinforce key concepts using notes features.</li> </ul>                                                                                                                                                      | <ul style="list-style-type: none"> <li>It has access to only limited number of E-Books.</li> <li>Impossible to upload whole E-Book.</li> </ul>                   |
| <b>Lumos Comprehend [30]</b>          | <ul style="list-style-type: none"> <li>An automated solution that helps to build quality questions and answers for textual content powered by advanced AI and ML algorithms.</li> <li>User can use this application to convert long articles into meaningful questions and answers.</li> <li>Once the questions and answers are generated, can view them on the screen or export them in CSV file format.</li> </ul> | <ul style="list-style-type: none"> <li>The system does not have text summarization functionality.</li> </ul>                                                     |

TABLE V. PERFORMANCE OF SYSTEM FOR QUESTION GENERATION

| No. of E-Books | Total Questions | Ideal Total Questions | Time required (minutes) | Relevant Questions | Irrelevant Questions | Percentage of Correct Questions | Percentage of Incorrect Questions |
|----------------|-----------------|-----------------------|-------------------------|--------------------|----------------------|---------------------------------|-----------------------------------|
| 5              | 159             | 200                   | 37.5                    | 126                | 33                   | 79.25                           | 20.75                             |
| 10             | 332             | 400                   | 77.5                    | 261                | 71                   | 78.61                           | 21.397                            |
| 15             | 469             | 600                   | 119                     | 372                | 97                   | 79.32                           | 20.68                             |
| 20             | 637             | 800                   | 160.5                   | 506                | 131                  | 79.43                           | 20.57                             |

TABLE VI. PERFORMANCE OF SYSTEM FOR DISTRACTOR GENERATION

| Num. of E-Books | Total MCQ Options Generated | Correct Options for MCQs | Incorrect Options for MCQs | Percentage of Relevant Options | Percentage of Irrelevant Options |
|-----------------|-----------------------------|--------------------------|----------------------------|--------------------------------|----------------------------------|
| 5               | 111                         | 82                       | 21                         | 73.87                          | 26.13                            |
| 10              | 246                         | 187                      | 39                         | 76.02                          | 23.98                            |
| 15              | 408                         | 319                      | 65                         | 78.19                          | 21.81                            |
| 20              | 537                         | 436                      | 77                         | 81.20                          | 18.81                            |

A collection of 20 sample E-Books were tested on the basis of above-mentioned sampling procedure on the system. The algorithms were trained to generate 8 questions per page. The Fig. 8 compares number of E-Books with time required to generate total questions. Questions generated by the system were considered to be relevant or irrelevant in accordance to grammatical and logical correctness in English language. For evaluation of the questions generated, it was found out that when 20 samples were tested, maximum accuracy of 79.435% is achieved. Otherwise for 5, 10 and 15 samples, accuracy of 79.245%, 78.614% and 79.317% is recorded respectively. The Fig. 9 compares ideal number of questions to be generated with the actual number of questions generated. The Fig. 10 depicts percentage relevancy of questions generated. For evaluation of the MCQs' options generated, it was observed that when 20 samples were tested, maximum accuracy of 81.199% is achieved. Otherwise for 5, 10 and 15 samples, accuracy of 73.873%, 76.016% and 78.186% is recorded, respectively. The Fig. 11 shows percentage relevancy of accurate options generated.

From the above data it can be summarized that with increased number of pages and question the accuracy of the

system went on increasing slightly. Also, the lower bound of the accuracy for relevant question generation was 78.614% and the same for relevant MCQ's option generation was 73.873%. Hence, we can conclude that the performance of the T5 transformer for question generation and question answering task, Distil BART for text summarization task and Sense2Vec for the distractor generation task was optimal and much better than the realm of mere guessing (50%).

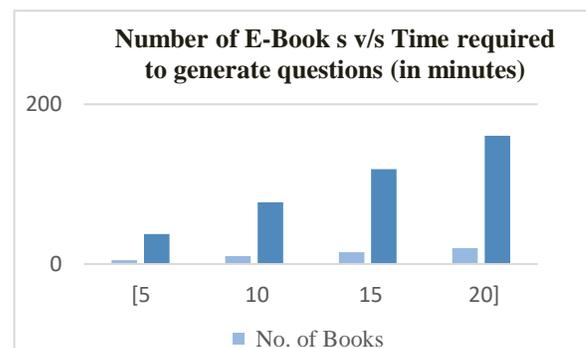


Fig. 8. Comparison of Number of E-Books and Time Required to generate Total Questions.

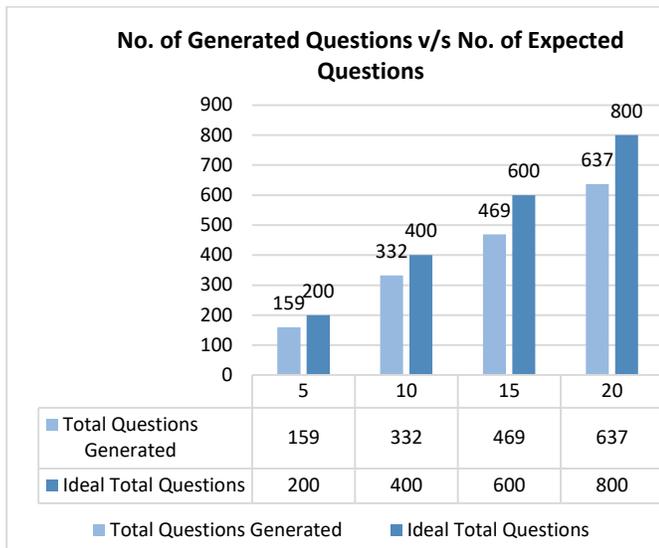


Fig. 9. Comparison of Ideal Number of Questions to be generated to Actual Number of Questions Generated.

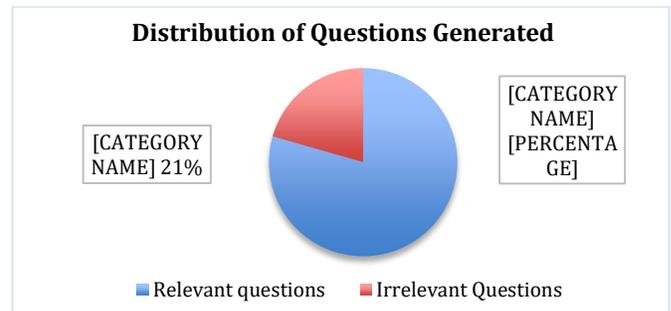


Fig. 10. Distribution of Questions Generated.

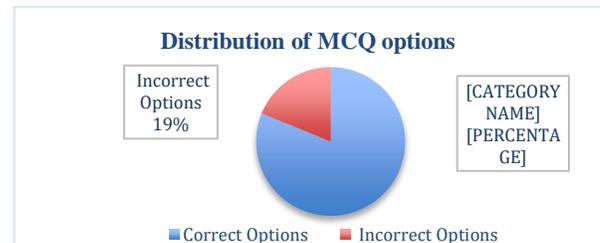


Fig. 11. Distribution of Multiple-Choice Questions.

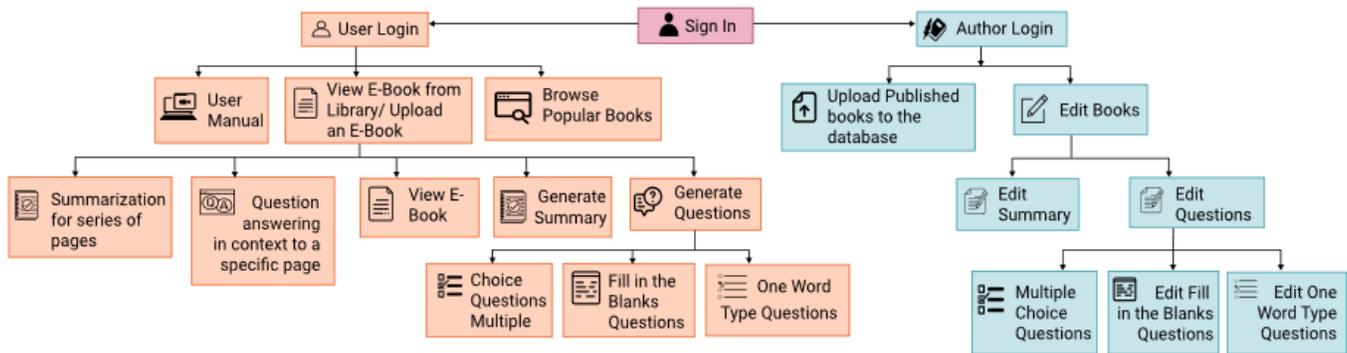


Fig. 12. Proposed System.

### IX. CONCLUSION

The boost in the amount of text data generated with time and development of technology has demanded research in automatic text summarization and question generation. Because of lockdown, E-reading and online examinations have turn out to be very popular, which includes many important examinations. No all-in-one system existed which provided both text summarization, question generation and question answering all at once. Hence, using NLP Transformers models like T5, Distil BERT, Distil BART this project creates such system resulting in reduced reading time and providing concise summary along with a questionnaire by implementing the existing algorithms with optimal accuracy. On implementing the fine-tuned transformers, efficient results are found out. Thus, the objectives for creating a system that provides a one stop destination solution to text summarization and question generation tasks were achieved.

For future work, the system can be elevated by scoring the readers on the basis of number of correct questions answered. This system can be used to evaluate the student’s capability and skills efficiently. Also, for upgradation of the system, focus will be on creating challenging questions for better learning

process. The system can be integrated with educational platforms like Moodle. A subscription model can also be created along with basic one. The paid version will have a feature that will enable users to communicate with the authors to solve queries. And authors will be updating the auto generated questions that they feel are semantically incorrect.

### REFERENCES

- [1] Edeh Michael Onyema, Chika Nwafor, et, al, “Impact of Coronavirus Pandemic on Education”, Journal of Education and Practice, Vol.11, No.13, 2020, pp 108-121, DOI: 10.7176/JEP/11-13-12.
- [2] N. Moratanch, S. Chitrakala, et al, "A survey on abstractive text summarization", 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT), March 2016, pp. 1-7, DOI: 10.1109/ICCPCT.2016.7530193.
- [3] Steven Bird, Ewan Klein, et al. O'Reilly Media, "Natural Language Processing with Python", Incorporated, 2009.
- [4] Sepp Hochreiter, Jurgen Schmidhuber, “Long Short-Term Memory”, PubMed, Neural Computation, 1997 Nov 15, pp. 1735-1780, DOI: 10.1162/neco.1997.9.8.1735.
- [5] Ashish Vaswani, Noam Shazeer, et, al, “Attention is All You Need”, 31st conference on Neural Information Processing Systems, NIPS 2017, Long Beach, CA, USA, pp. 6000-6010, arXiv:1706.03762.
- [6] Albert Zeyer, Parnia Bahar, et, al, "A Comparison of Transformer and LSTM Encoder Decoder Models for ASR", 2019 IEEE Automatic

- Speech Recognition and Understanding Workshop (ASRU), Singapore, December 2019, pp. 8-15, DOI: 10.1109/ASRU46091.2019.9004025.
- [7] Colin Raffel, Noam Shazeer, et, al, "Exploring the Limits of Transfer Learning with Unified Text-to-Text Transformer", Journal of Machine Learning Research 21(2020), June 2020, arXiv:1910.10683.
- [8] Jacob Devlin, Ming-Wei Chang, et, al, "BERT: Pre-Training of Deep Bidirectional Transformer for Language Understanding", Google AI Language, May 2019, arXiv:1810.04805.
- [9] Mike Lewis, Yinhan Liu, et, al, "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation and Comprehension", Facebook AI, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 2021, arXiv:1910.13461.
- [10] Princeton University, "WordNet-A Lexical Database for English", Accessed 01 August 2021, <https://wordnet.princeton.edu/>.
- [11] Julie Weeds, Daoud Clarke, et. al, "Learning to Distinguish Hypernyms and Co-Hyponyms", Proceeding of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pp. 2249-2259, August 2014.
- [12] Yuni Susanti, Takenobu Tokunaga, et.al, "Automatic Distractor generation For Multiple Choice English Vocabulary Question", Research and Practice in Technology Enhanced Learning 13, Article Number: 15, October 2018, DOI: 10.1186/s41039-018-0082-z.
- [13] Hugo Liu, Push Singh, et. Al, "The ConceptNet Project V2.1", Accessed July 2021, <http://alumni.media.mit.edu/~hugo/conceptnest/#papers>.
- [14] Ke Shen, Mayank Kejriwal, "A Data-Driven Study of Common-sense Knowledge Using the ConceptNet Knowledge Base", Jan 2021, arXiv:2011.14084.
- [15] MongoDB, "Welcome to the MongoDB Documentation", Accessed July 2021, <https://docs.mongodb.com/>.
- [16] SpaCy, "Library Architecture", Accessed July 2021, <https://spacy.io/api>.
- [17] Jaime Carbonell, Jade Goldstein, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries", SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR conference on Research and Development in Information retrieval, pp. 335-336, Issue August 1998, DOI: 10.1145/290941.291025.
- [18] Mengyu Li, Boyao Sun, et, al, "Question Answering on SQuAD2.0", in Stanford CS224N Natural Language Processing with Deep Learning.
- [19] Stuart rose, Dave Engel, et, al, "Automatic Keyword Extraction from Individual Documents", in Text Mining: Application and Theory, pp. 1-20, Issue March 2010, DOI: 10.1002/9780470689646.ch1.
- [20] Google Firebase. "Google Firebase", Accessed July 2021, [https://firebase.google.com/?gclid=Cj0kCQjw24qHBhCnARIsAPbdtl3nmCf5-9fIE3SiyvTqT3cjxcQ7piDulL\\_j24QDLarNu5YgUeJLkaAiJREALw\\_wcB&gclsrc=aw.ds](https://firebase.google.com/?gclid=Cj0kCQjw24qHBhCnARIsAPbdtl3nmCf5-9fIE3SiyvTqT3cjxcQ7piDulL_j24QDLarNu5YgUeJLkaAiJREALw_wcB&gclsrc=aw.ds).
- [21] Bojar et al. "WMT 2016", Findings of the 2016 Conference on Machine Translation. Accessed September 2021, <https://paperswithcode.com/dataset/wmt-2016>.
- [22] Bojar, Rajen Chatterjee, et, al, "Findings of the 2016 Conference on Machine Translation (WMT-2016)", Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, Berlin, Germany, pp. 131-198, Issue August 2016, DOI: 10.18653/v1/W16-2301.
- [23] Ramesh Nallapati, Bowen Zhou et, al, "Abstractive text Summarization Using Sequence-to-Sequence RNNs and Beyond", CNN/Daily Mail. Accessed September 2021, <https://paperswithcode.com/dataset/cnn-daily-mail-1>.
- [24] Chin-Yew Lin. "Looking for a few Good Metrics: ROUGE and its Evaluation", Working Notes of NTCIR-4, Tokyo, Issue June 2004, Corpus ID: 55156862.
- [25] Alex Wang, Amanpreet Singh, et. Al, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding", Issue April 2018, arXiv:1804.07461.
- [26] Perer Flach, Meelis Kull, "Precision-Recall-Gain Curves: PR Analysis Done Right" In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 28, Vol. 1, pp. 838-846, DOI: 10.1.1.1038.8511.
- [27] Text Summarization | Text Summarizer, "Text Summarization", Accessed July 2021, <http://textsummarization.net/>.
- [28] AutoSummarizer.com, "Automatic Text Summarizer", Accessed July 2021, <https://autosummarizer.com/>.
- [29] Harbinger AI Inc, "Quillionz", Accessed July 2021, <https://www.quillionz.com>.
- [30] Lumos Learning, "Lumos Comprehend", Accessed July 2021, <https://www.lumoslearning.com/llwp/free-question-answer-generator-online.html>.

# A Regression Model to Predict Key Performance Indicators in Higher Education Enrollments

Ashraf Abdelhadi, Suhaila Zainudin, Nor Samsiah Sani  
Center for Artificial Intelligence Technology (CAIT)  
Faculty of Information Science and Technology  
Universiti Kebangsaan Malaysia  
Bangi, Selangor, Malaysia

**Abstract**—Key Performance Indicators (KPIs) are essential factors for the success of an organization. KPIs measure the current performance and identify the ongoing progress for specified business objectives. The Ministry of Higher Education (MoHE) in Palestine used established formulas to predict the KPI. These KPIs are vital for charting the organization aims. This study applies regression models for student enrollment data sets to predict accurate KPIs that can be used and adapted for any higher education system. The predictive engine will determine the KPI based on linear regression techniques such as Lasso, Elastic Net, and non-linear regression such as Support Vector Regression (SVR), and K-Nearest Neighbor (KNN). The Ministry of Higher Education (MoHE) in Palestine provided the datasets related to enrollments and graduations data for different Higher Education Institutions (HEIs). The regression algorithms were evaluated by mean absolute error, mean square error (MSE), root mean square error (RMSE) and the R Squared. The experiment demonstrates that the 40% training with 60% testing splitting using linear regression shows the best result.

**Keywords**—Data mining; KPI; regression; higher education; prediction model

## I. INTRODUCTION

Key Performance Indicators (KPIs) are the critical signs of development in the direction of a meant result. KPIs afford a focal point for strategic and operational improvement, create an analytical foundation for decision-making, and assist awareness interest on most topics. KPI performs a critical element given that it is given fast and specific data through evaluating present-day overall performance in opposition to a goal required to fulfil commercial enterprise desires and objectives [1].

Businesses adopted frameworks such as the balanced scorecard (BSC) [2] as a strategic performance metric to improve internal business functions and their outcomes. Correspondingly, education centers, knowledge creation and worker centers such as ministries or learning institutions also benefited from utilizing BSC to chart the KPIs for Higher Education Institutions [3]. On a global scale, the United Nations Educational, Scientific, and Cultural Organization (UNESCO) published a practical guide for educational planners who wish to construct an indicator system [4]. The author in [4] included examples of HEIs, notably the University of Edinburgh and University Technology Malaysia, that planned their strategic development plans alongside a monitoring system, such as BSC.

The structure and content of education systems around the world vary greatly. As a result, they compare national education systems with other countries or benchmark progress toward national and international goals. Hence, UNESCO designed the International Standard Classification of Education (ISCED) to serve as a framework to classify educational activities as defined in programs and the resulting qualifications into internationally agreed categories. The basic concepts and definitions of ISCED are internationally valid and comprehensive of the full range of education systems [5].

However, there is no solid data mining framework and model to predict the higher education (HE) KPI across the world and at MoHE Palestine in particular. For instance, the current MoHE practice to extract and predict KPI is manual. The staff collect the data from different resources by phone and emails, then record it into an excel sheet, as shown in Fig. 1. The formula miscalculated will lead to a wrong decision.

| Indicator definition                                                                                                                        | Indicator formula (calculation)                                                                                                            |
|---------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|
| All students enrolled in post-secondary educational institutions in relation to the to the age category of tertiary education in Palestine. | $\frac{\text{Number of enrolled students for the academic year}}{\text{Number of population age category 18 - 33 in the same year}}$       |
| All enrolled students in post-secondary educational institutions, distributed according to level of study (ISCED5-8)                        | $\frac{\text{Number of enrolled students in a specific level}}{\text{The overall number of enrolled students}}$                            |
| All enrolled students in post-secondary educational institutions, distributed according to level of study (ISCED9)in Jordan                 | $\frac{\text{Number of students in short - cycle tertiary education (level 5)}}{\text{Number of students in tertiary education}}$          |
| All enrolled students in post-secondary educational institutions, distributed according to level of study (ISCED9)in India                  | $\frac{\text{Number of students in Bachelor or equivalent tertiary education (level 6)}}{\text{Number of students in tertiary education}}$ |
| All enrolled students in post-secondary educational institutions, distributed according to level of study (ISCED7)in Mexico                 | $\frac{\text{Number of students in master or equivalent tertiary education (level 7)}}{\text{Number of students in tertiary education}}$   |

Fig. 1. Example of KPIs Formulation.

## II. PROBLEM STATEMENT

Although the MoHE has computerized most of its services and automated most operations, the ministry is still facing some issues in the reporting system and predication, which affects the strategic plan for the upcoming years, for instance, predicting wrong enrollment students' number for the forthcoming academic year in government Tertiary Education Institutes (TEIs) can cause in improper budget allocation which means wasting of resources. Also, extracting knowledge from complex data sets takes a long time and a human effort to drill deeply into the big data sets. Therefore, the main worthwhile problem that needs to be addressed is to discover a new fast, efficient, and incredibly accurate

computerized approach or data mining algorithm to resolve the KPI extraction and prediction problem primarily for our case study (MoHE) based on the database for the benefit of the higher education management.

Data availability, especially for the education sector, has spurred interest in data-driven decision making [6]. The process of making organizational decisions based on actual data rather than intuition or observation alone is known as data-driven decision making (or DDDM). Therefore, DDDM offers the opportunity to discover a new fast, efficient, and incredibly accurate computerized approach or data mining algorithm to resolve KPI extraction and prediction for the MoHE case study.

### III. RELATED WORK

Data mining includes many techniques from other domains such as statistics, machine learning, pattern recognition, database, data warehouse systems and visualization [7]. Most organizations monitor their operation performance and achievement through dashboards and Business Intelligence (BI) [8]. However, in many institutes, this is limited to standard reports which cannot measure the unknown KPIs and in most cases, it is difficult to predict future performance. Most top managers rely on their intuition in order to select their potential KPIs that will lead to redundant KPIs. Managers also focus on the results rather than on the actual indicators that can be used [1].

The author in [1] built a model to predict key performance indicators for Massive Online Open Courses (MOOC) that is very similar to the Cross-Industry Standard Process for Data Mining (CRISP-DM). The model consisted of six stages from defining the business strategy model, definition of KPIs and the multidimensional model. The multidimensional model is composed of two analysis cubes: Enrollment and Activity. The enrollment analyzes the students' features such as country, interests and expectations and whether these features represent specific patterns. Data mining techniques are used to extract and predict KPIs. These techniques analyze the KPIs to mine the relationships identified during the business strategy modelling. The author in [1] used different algorithms such as Support Vector Machines (SVM), a Random Forest of Decision Trees (DT) and Neural Networks.

In 2015, [9] proposed a framework for predicting students' academic performance. The primary purpose is to discover hidden information and knowledge from the students' data so that the model can predict the student grades in a specific subject based on independent parameters such as GPA, race, gender, family income, university entry mode. The model proposed in [9] used three different classifications algorithms: Decision Tree (DT), Naïve Bayes (NB), and Rule-Based (RB) through the WEKA software tool. The model allows users to categorize the students under two or three categories; good, poor, and average. If this framework and model can be modified to use a regression algorithm, the output can be numbers or percentages, which is more accurate.

The author in [10] built a model to classify attrition among B40 students in bachelor's degree programs in Malaysia's public universities. The machine learning model indicates that

the Random Forest algorithm is the best model in predicting student attrition compared to Neural Network and Decision Tree.

The author in [11] applied different machine learning techniques to qualitatively predict the whole project KPIs in critical construction project stages. Artificial neural network (ANN) and the neuro-fuzzy method using fuzzy C-means (FCM) and subtractive clustering to predict project KPIs. The models map the KPIs of three critical project stages to the whole project KPIs. Validation used the data of actual projects to confirm models' effectiveness and compare the results of the employed machine learning techniques.

The author in [12] created a model to predict and identify factors that influence graduates' employability. Seven years of data (from 2011 to 2017) from Malaysia's Ministry of Education were used to test and evaluate the model. They applied three different algorithms; Decision Tree, Support Vector Machines and Artificial Neural Networks. The results show the decision tree (J48) produces higher accuracy compared to other techniques. Also, according to this study, three factors, attribute age, industrial internship, and faculty, contain the most information and affect the final class, which is employability.

TABLE I. SUMMARY OF RELATED WORK

| Reference | Theme (concept)                                | Findings/Conclusions                                                                                                                                                                                                                                                                                                                                                                 |
|-----------|------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [1]       | Data mining framework and KPI Predictive model | It is a good model but without a clear framework that can cover the whole KPIs prediction process.                                                                                                                                                                                                                                                                                   |
| [9]       | Data mining framework and KPI Predictive model | The model can predict the student grades (dependent parameter) in a specific subject based on independent parameters such as GPA, race, gender, family income, university entry mode. The study focused on being more comparative between three algorithms. The result is a lack of graphs and charts that clearly show the output and the output discrete, not a continuous number. |
| [11]      | KPI Predictive model                           | All KPIs were measured qualitatively by designing a questionnaire, and there is no database containing accurate records. Also, The research measures project performance from the owner's point of view.                                                                                                                                                                             |
| [12]      | Predictive model                               | Created a model to predict and identify factors that influence graduates' employability.                                                                                                                                                                                                                                                                                             |
| [10]      | Predictive model                               | Built a model to classify attrition among B40 students in bachelor's degree programs in Malaysia's public universities.                                                                                                                                                                                                                                                              |

### IV. MATERIALS AND METHODS

The Cross-Industry Standard Process for Data Mining (CRISP-DM) is a methodology model with six stages describing the information technology existence cycle. It will help plan, organize, and enforce data science (or machine learning) tasks Fig. 2. It standardizes data mining techniques throughout industries, analytics, and data science projects.

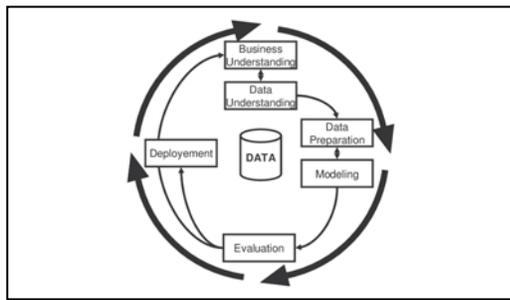


Fig. 2. CRISP-DM Diagram.

The six CRISP-DM Phases are Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

The research experiment will apply the six CRISP-DM Phases on the Ministry of Higher Education and Scientific Research (MoHE) in Palestine [13] focus on the KPIs related to students' enrollments according to INTERNATIONAL STANDARD CLASSIFICATION OF EDUCATION (ISCED) [14] such as enrolled students in post-secondary educational institutions, distributed according to each field of study:

- 1) According to Education Program.
- 2) Arts and Humanities.
- 3) Social Science, Journalism and information.
- 4) Business, Administration and law.
- 5) Natural Science, Math and Science.
- 6) Information and communication technology.
- 7) Engineering Manufacturing and Constructions.
- 8) Agriculture, Forestry, Fisheries and Veterinary.
- 9) Health and Welfare.
- 10) Services.

UNESCO designs ISCED to serve as a framework to classify educational activities as defined in programs and the resulting qualifications into internationally agreed categories. Therefore, the basic concepts and definitions of ISCED are intended to be internationally valid and comprehensive of the full range of education systems [14].

The details of the methodology followed in this study is explained below.

#### Phase 1: Business Understanding

This phase concerns determining the business goals which is to predict a set of KPIs that has been defined in higher education and the best practice to measure those KPIs.

#### Phase 2: Data Understanding and Data Resources Analysis

At this phase, the data resources have been prepared for modelling, including several activities such as data selection, data cleaning, data construction, data integration, combining data from multiple sources, and re-formatting data as necessary. Data Source identification (databases, schema names, tables, view, spreadsheets), SQL scripts performed to create specific views in the staging database, combine data from multiple sources to one repository pre-processing data stage, including data selection, cleaning and integration.

The enrollment and graduation data form the core sources [13] for our data mining experiments. For instance, the original enrollment table consists of 50 attributes and 3,895,158 instances as it contains the historical data since the MoHE establishment. The enrollment attributes (fields) were identified to contain 34 features and 3862763 instances as some fields duplicated for both English and Arabic values. The graduation data sets have 461,598 instances and 24 attributes.

#### Building Database Repository using SQL server:

The database repository is built based on main tables such as enrollment, graduations, ISCED levels, programs, and degrees, in addition to many lookup tables such as high schools' lists, districts, nationalities, universities lists. Data views were created to focus on the data from 2014 to 2018, including 25 attributes and other attributes from other tables containing the ISCED data, which is essential for data mining. Some repeated features (fields) such as the Arabic values have been eliminated because it's considered duplicate values, other values replaced with null values excluded.

#### Data Cleaning and Transformation:

Any noisy and inconsistent data were removed to handle the missing data fields, transform data into forms appropriate for the mining task, for instance, the area code to numbers from 1 to 16, the high school types coded from 1 to 5 and the high school stream coded to numbers as well (Tables II, III and IV) The data is split into 60% training and 40% testing sets.

TABLE II. AREA CODE DATA TRANSFORMATION

| CODE | Area        |
|------|-------------|
| 1    | Quds        |
| 2    | Hebron      |
| 3    | Ramallah    |
| 4    | Bethlehem   |
| 5    | Nablus      |
| 6    | Tukaram     |
| 7    | Qalqilya    |
| 8    | Sal fit     |
| 9    | Jenin       |
| 10   | Jericho     |
| 11   | Gaza        |
| 12   | Middle Gaza |
| 13   | Khan Younis |
| 14   | DerAlbalah  |
| 15   | Rafah       |
| 16   | Tubas       |

TABLE III. HIGH SCHOOL TYPE DATA TRANSFORMATION

| CODE | HS_Type          |
|------|------------------|
| 1    | Gov. High School |
| 2    | Bajrout          |
| 3    | GCE              |
| 4    | IB               |
| 5    | SAT              |

TABLE IV. HIGH SCHOOL STREAM DATA TRANSFORMATION

| CODE | HS-Stream           |
|------|---------------------|
| 1    | Humanities          |
| 2    | Literature          |
| 3    | Science             |
| 4    | Industry            |
| 5    | Economic            |
| 6    | Agriculture         |
| 7    | Nursing             |
| 8    | Hospitality         |
| 9    | Islamic Study       |
| 13   | Applied Industry    |
| 14   | Applied Agriculture |
| 15   | Vocational          |
| 19   | IT                  |
| 20   | Entrepreneurship    |
| 21   | Technology          |

### Phase 3: Modeling

Regression predicts a range of numeric values or continuous values. For example, a regression model that predicts KPI values could be developed based on observed data for many other factors such as enrolled students, specific programs, number of graduates throughout history.

Numerous models were constructed and assessed primarily based on numerous techniques. In this study employed Linear Algorithms: Linear Regression (LR), Lasso Regression (LASSO) and Elastic Net. The study also applied nonlinear algorithms such as Support Vector Regression (SVR), and K-Nearest Neighbors (KNN) using Python. In terms of parametrization, the variable "ISCED\_Level1\_Id" is assigned as a target to be predicted. To generate the training, the random\_state variable is assigned to 1 to replicate results with frac=0.6. Then select any data, not in the training set and include it in the testing set based on the index, test = df.loc [~df.index.isin(train.index)].

### Predicting ISCED KPIs:

In this study, the first experiment is to predict the KPIs for enrolled students in post-secondary educational institutions, distributed according to every ten fields of study based on the first level of (ISCED) and the general studies. So, the model will predict KPIs according to the 10 identified field of study. The second experiment is to predicts find the ratio between enrollment and graduation based on the graduates data sets.

### Phase 4: Evaluation

There are three metrics for evaluating predictions in regression; Mean Absolute Error, Mean Squared Error, and R2. The Mean Absolute Error (or MAE) is the sum of the absolute differences between predictions and actual values. It gives an idea of how wrong the predictions were. The measure provides a picture of the magnitude of the error but no idea of the direction (e.g., over or under predicting). A value of 0 indicates no error or perfect predictions.

The Mean Squared Error (or MSE) is just like the implied absolute mistakes in that it affords a gross concept of the significance of the mistakes. Taking the rectangular root of the implied squared mistakes converts the units lower back to the unique units of the output variable and may be significant for description and presentation. This is referred to as the Root Mean Squared Error (or RMSE). So, for instance, if MSE= - 34.705 and SD =45.574, this metric is inverted to increase the outcomes.

The R2 (or R Squared) metric illustrates the goodness in the shape of a fixed of predictions to the actual values. In statistical literature, this degree is referred to as the coefficient of determination. This is a value among zero and 1 for no-match and best match, respectively. For example, if R2 = 0.2, the predictions have a negative match to the real values with a value toward 0 and much less than 0.5. The last stage is the deployment with the task of plan deployment and tracking, produce the final report, and review tasks by conducting an assignment retrospective approximately what went well, what might have been better, and a way to enhance it [1].

## V. EXPERIMENTAL RESULTS AND DISCUSSION

Before applying different algorithms for different datasets based on the academic years, the three linear regression algorithms were tested with varying percentages of splitting of training and testing data (10% to 90%) (Table V).

According to [15] (scikit-learn, 2021), Linear Regression fits a linear model with coefficients  $w = (w_1, \dots, w_p)$  to minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation. Mathematically it solves a problem of the form;

$$\min_w ||Xw - y||_2^2$$

TABLE V. SHOWS LINEAR REGRESSION RESULTS FOR DIFFERENT SPLIT

| Algorithm Linear Regression | Time (S)  | MAE                     | MSE                | RMSE               | R Squared       |
|-----------------------------|-----------|-------------------------|--------------------|--------------------|-----------------|
| 10% to 90%                  | 9.6       | 0.000000062<br>10088990 | 0.00000<br>099729  | 0.00099864<br>4080 | 0.987710<br>90  |
| 20% to 80%                  | 9.95      | 0.000000062<br>20099000 | 0.00000<br>500290  | 0.00223671<br>6340 | 0.978710<br>90  |
| 30% to 70%                  | 9.95      | 0.000000050<br>18788899 | 0.00000<br>149280  | 0.00122180<br>1959 | 0.968098<br>0   |
| 40% to 60%                  | 9.87      | 0.000000042<br>18000023 | 0.00000<br>098129  | 0.00099060<br>0837 | 0.998719<br>90  |
| 50% to 50%                  | 9.9       | 0.000000076<br>57778899 | 0.00000<br>145280  | 0.00120532<br>1530 | 0.977714<br>40  |
| 60% to 40%                  | 10.4<br>9 | 0.000000046<br>56890001 | 0.00000<br>0997522 | 0.00099876<br>1990 | 0.988710<br>90  |
| 70% to 30%                  | 9.95      | 0.000000058<br>65412345 | 0.00000<br>090020  | 0.00094878<br>8709 | 0.977087<br>155 |
| 80% to 20%                  | 11.2<br>1 | 0.000000066<br>43210008 | 0.00000<br>172280  | 0.00131255<br>4765 | 0.955718<br>70  |
| 90% to 10%                  | 10.1<br>5 | 0.000000080<br>90087799 | 0.00000<br>242280  | 0.00155653<br>4613 | 0.908710<br>90  |

Lasso Regression:

The Lasso is a linear model that estimates sparse coefficients. It is helpful to prefer solutions with fewer non-zero coefficients, effectively reducing the number of features the given answer depends on. For this reason, Lasso and its variants are fundamental to the field of compressed sensing. Under certain conditions, it can recover the exact set of non-zero coefficients.

$$\min_w \frac{1}{2n_{\text{samples}}} \|Xw - y\|_2^2 + \alpha \|w\|_1$$

The Lasso estimate thus solves the minimization of the least-squares penalty with  $\alpha \|w\|_1$  is the  $l_1$  The implementation in the class Lasso uses coordinate descent as the algorithm to fit the coefficients [15] (scikit-learn, 2021). Table VI shows the experimental results for lasso regression when applying different splitting.

Elastic Net is a linear regression model trained with both  $l_1$  and  $l_2$ -norm regularization of the coefficients. This combination allows for learning a sparse model where few of the weights are non-zero, like Lasso Elastic-net, which is beneficial for multiple features correlated with each other, such as high school average and high school stream. Lasso is likely to pick one of these at random, while elastic-net is likely to determine both [15] (scikit-learn, 2021).

$$\min_w \frac{1}{2n_{\text{samples}}} \|Xw - y\|_2^2 + \alpha \rho \|w\|_1 + \frac{\alpha(1 - \rho)}{2} \|w\|_2^2$$

Table VII shows the experimental results for Elastic Net regression when applying different splitting.

When comparing the three linear algorithms (Fig. 3), the Linear algorithm score the lowest error compared to the Lasso and Elastic Net.

TABLE VI. LASSO REGRESSION FOR DIFFERENT SPLITTING

| Algorithm (Lasso Regression) | Time (S) | MAE                | MSE                | RMSE             | R Squared          |
|------------------------------|----------|--------------------|--------------------|------------------|--------------------|
| 10% to 90%                   | 9.65     | 0.227520<br>37948  | 0.12489<br>1937503 | 0.3534005<br>341 | 0.98283669<br>9366 |
| 20% to 80%                   | 9.73     | 0.228279<br>447022 | 0.12759<br>1253438 | 0.3571991<br>789 | 0.98248180<br>2987 |
| 30% to 70%                   | 9.47     | 0.227984<br>559434 | 0.12576<br>3102058 | 0.3546309<br>378 | 0.98273225<br>6546 |
| 40% to 60%                   | 9.65     | 0.220175<br>387930 | 0.12621<br>0860621 | 0.3552616<br>790 | 0.98964821<br>3930 |
| 50% to 50%                   | 9.2      | 0.228149<br>889028 | 0.12697<br>3529679 | 0.3563334<br>529 | 0.98252568<br>0214 |
| 60% to 40%                   | 9.72     | 0.227763<br>064940 | 0.12623<br>2959929 | 0.3552927<br>805 | 0.98265989<br>2778 |
| 70% to 30%                   | 9.43     | 0.226355<br>016909 | 0.12572<br>7079609 | 0.3545801<br>455 | 0.98268306<br>3141 |
| 20% to 80%                   | 9.8      | 0.229283<br>720738 | 0.12832<br>0218693 | 0.3582181<br>160 | 0.98230214<br>3850 |
| 90% to 10%                   | 9.7      | 0.229283<br>720738 | 0.12832<br>0218693 | 0.3582181<br>160 | 0.98230214<br>3850 |

TABLE VII. ELASTIC NET REGRESSION FOR DIFFERENT SPLITTING

| Algorithm (Elastic Net Reg.) | Time (S)  | MAE                       | MSE                   | RMSE                      | R Squared             |
|------------------------------|-----------|---------------------------|-----------------------|---------------------------|-----------------------|
| 10% to 90%                   | 11        | 0.228353<br>9916884<br>26 | 0.12451944<br>8633917 | 0.352873<br>1339078<br>07 | 0.98311322<br>1419126 |
| 20% to 80%                   | 10        | 0.228482<br>1395454<br>30 | 0.12488748<br>5646334 | 0.353394<br>2354458<br>18 | 0.98312134<br>3318927 |
| 30% to 70%                   | 12        | 0.228194<br>4875470<br>74 | 0.12531987<br>7380630 | 0.354005<br>4764839<br>52 | 0.98310014<br>7069539 |
| 40% to 60%                   | 9.43      | 0.221302<br>6070346<br>49 | 0.12498577<br>6002300 | 0.353533<br>2742505<br>30 | 0.98900381<br>4519843 |
| 50% to 50%                   | 11.3<br>5 | 0.228598<br>5855036<br>90 | 0.12486465<br>0076610 | 0.353361<br>9250522<br>20 | 0.98305375<br>8064118 |
| 60% to 40%                   | 12        | 0.228957<br>8469231<br>30 | 0.12494810<br>5589363 | 0.353479<br>9931953<br>19 | 0.98318187<br>2293709 |
| 70% to 30%                   | 11.7<br>7 | 0.228222<br>9513866<br>59 | 0.12347448<br>4082960 | 0.351389<br>3625068<br>35 | 0.98337124<br>158077  |
| 80% to 20%                   | 9.45      | 0.229529<br>0235438<br>00 | 0.12821437<br>0176883 | 0.358070<br>3424983<br>46 | 0.98272818<br>4185912 |
| 90% to 10%                   | 9.65      | 0.228076<br>2294707<br>95 | 0.12005582<br>5721761 | 0.346490<br>7296332<br>19 | 0.98359149<br>0829027 |

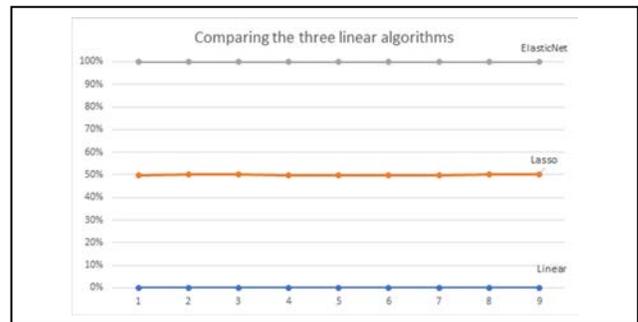


Fig. 3. Comparison between the three different Linear Algorithms.

Moreover, the slightest error margin was 40% training and 60% testing data sets (Fig. 4). Therefore, the rest of the algorithms tested for the exact sampling percentages (40% training and 60% testing) for the same academic year. Then, we look at the different iterations for three algorithms with varying percentages of data sampling (training and testing). There is no significant difference using the same algorithm for further selection, but there is a difference when it comes to the non-linear algorithms.

The experiment conducted for the same academic year, the enrollment KPI was based on ISCED level one for five different algorithms, as shown in Table VIII.

The ISCED KPIs predicted values for enrolled students in post-secondary educational institutions, distributed according to the 10 fields of study. Fig. 5 shows that the values are very close to the actual values.

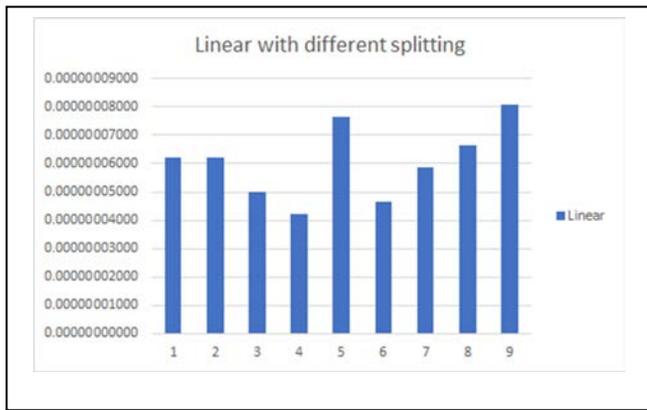


Fig. 4. Margin Error based on % Splitting Sampling.

TABLE VIII. KPI PREDICTION ERROR

| Algorithm                       | Time(M. S.MS) | MAE        | MSE               | RMSE         | R2           |
|---------------------------------|---------------|------------|-------------------|--------------|--------------|
| Linear Regression               | 9.87          | 0.0000004  | 0.000000981       | 0.000990454  | 0.99871090   |
| Lasso Regression                | 9.65          | 0.22752037 | 0.123148555719873 | 0.3509252850 | 0.9831577460 |
| Elastic Net Regression          | 9.43          | 0.22887746 | 0.125000076102056 | 0.3535534982 | 0.9830981775 |
| Support Vector Regression (SVR) | 47.71         | 0.66592015 | 6.968257047       | 2.6397456407 | -16.320      |
| K-Nearest Neighbors (KNN)       | 3.29.50       | 0.66537128 | 1.406274302       | 1.1858643694 | 0.7453734924 |

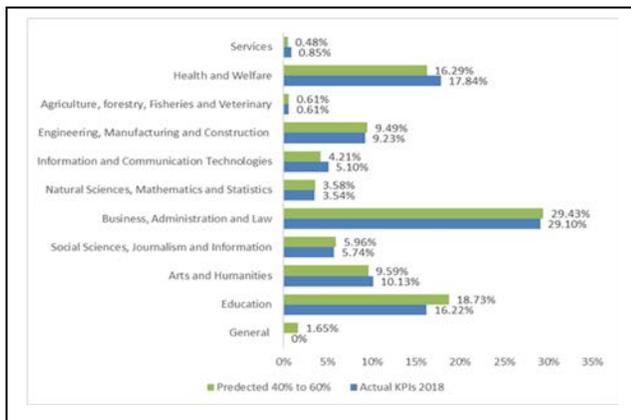


Fig. 5. Comparison between the Actual and Predicted ISCED KPIs.

The second Experiment was to find predicted ratio between enrollment and graduation. Fig. 6 shows the ratio between the predicted enrollment and graduation KPIs based on ISCED level 1.

The ISCED numbers in Fig. 6 can be translated as per Table IX.

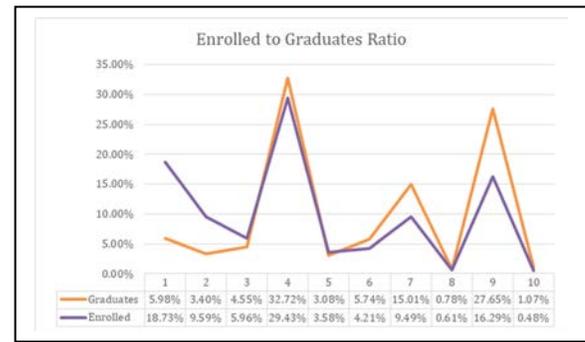


Fig. 6. Enrollment to Graduates Ratio.

TABLE IX. ISCED LEVEL 1 DESCRIPTION

| ISCED_Level1_Description                        | ISCED_Level1_Id |
|-------------------------------------------------|-----------------|
| Education                                       | 1               |
| Arts and Humanities                             | 2               |
| Social Sciences, Journalism and Information     | 3               |
| Business, Administration and Law                | 4               |
| Natural Sciences, Mathematics and Statistics    | 5               |
| Information and Communication Technologies      | 6               |
| Engineering, Manufacturing and Construction     | 7               |
| Agriculture, forestry, Fisheries and Veterinary | 8               |
| Health and Welfare                              | 9               |
| Services                                        | 10              |

## VI. CONCLUSION

In conclusion, with clear and coherent strategies, figuring out the present-day situations, operation sector, unique varieties of competencies that generate, performance will lead to success. To create this kind of situation calls for the provision of strategic records to confirm the current situations, to outline the strategy [16]. Also, applying Cross-Industry Standard Process for Data Mining (CRISP-DM) process model as a research methodology to develop a data mining model that could help be adapted by individuals and HEIs, using machine learning algorithms can lead to good results and accuracy [17]. However, without clear KPIs, it's challenging to have a clear strategy for the upcoming years. It is crucial to create an analytical model to act as the basis for decision making and help focus attention on HE enrollment. This study provides a practical solution for such a problem by proposing a KPIs predicting model from available data at MoHE and integrating the data from different resources into a database repository from which KPIs will be predicted. This model tested different regression algorithms such as linear regression, Lasso, Elastic Net; non-linear Support Vector Regression (SVR) and K-Nearest Neighbors (KNN). However, the most successful predictive model and particularly in performance indicators used was Linear regression. The training and splitting data were tested from 10% to 90%, the targets values were compared from the historical data in the last few years. The regression algorithms were evaluated by mean absolute error, mean square error (MSE), root mean square error (RMSE) and the R Squared. The 40% training

with 60% testing splitting using linear regression shows the best result. In the future, this model can be part of a complete HE Framework to predict the KPIs and act as the main engine for that Framework.

#### ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) and DTK Grant (TT-2020-015) for supporting this research and the General Directorate of Research and Development at Ministry of Higher Education and Scientific research-Palestine for providing the data.

#### REFERENCES

- [1] Peral, J., Maté, A. and Marco M. Application of Data Mining techniques to identify relevant Key Performance Indicators. *Computer Standards and Interfaces Volume 54, Part 2, November 2017, Pages 76-85, 2017.*
- [2] Kaplan, R. S. 2009. Conceptual Foundations of the Balanced Scorecard. *Handbooks of Management Accounting Research 3: 1253–1269. doi:10.1016/S1751-3243(07)03003-9.*
- [3] Weerasooriya, R. B. Adoption of the Balanced Scorecard (BSC) Framework as a Technique for Performance Evaluation in Sri Lankan Universities. *SSRN Electronic Journal (November). doi:10.2139/ssrn.2223933, 2013.*
- [4] Martin, M. and Sauvageot, C. 2011. Constructing an indicator system or scorecard for higher educ Martin, M. and Sauvageot, C. *Constructing an Indicator System or Scorecard for Higher Education. A Practical Guide. UNESCO International Institute for Educational Planning. Paris, 2011. ISBN: 978-92-803-1329-1. Pages: 83.*
- [5] I. S. The International Standard Classification of Education (ISCED). In *Prospects (Vol. 5, Issue 2), 1975. [online]. Available: https://doi.org/10.1007/BF02207511.*
- [6] Ballou, Brian, Heitger, Dan L. and Stoel, Dale, (2018), Data-driven decision-making and its impact on accounting undergraduate curriculum, *Journal of Accounting Education, 44, issue C, p. 14-24, https://EconPapers.repec.org/RePEc:eee:joaced:v:44:y:2018:i:c:p:14-24.*
- [7] Hartama, D., Windarto, A. P., and Wanto, A. (2019). The application of data mining in determining patterns of interest of high school graduates. *Journal of Physics: Conference Series, 1339(1) doi:http://dx.doi.org/10.1088/1742-6596/1339/1/012042[8]* Stefanovic, N. 2015. Collaborative predictive business intelligence model for spare parts inventory replenishment. *Computer Science and Information Systems 12(3): 911–930. doi:10.2298/CSIS141101034S.*
- [8] Ahmad, F., Ismail, N. H. and Aziz, A. A. The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences 9(129): 6415–6426. doi:10.12988/ams.2015.53289, 2015.*
- [9] Sani, N. S., Nafuri, A. F. M., Othman, Z. A., Nazri, M. Z. A., and Nadiyah Mohamad, K. Dropout Prediction in Higher Education Among B40 Students. *International Journal of Advanced Computer Science and Applications, 11(11), 550–559. https://doi.org/10.14569/IJACSA.2020.01111169, 2020.*
- [10] Fanaei, S. S., Moselhi, O., Alkass, S. T. and Zangenehmadar, Z. Application of Machine Learning in Predicting Key Performance Indicators for Construction Projects. *International Research Journal of Engineering and Technology: 1450. Retrieved from www.irjet.net, 2018.*
- [11] Othman, Z., Shan, S. W., Yusoff, I., and Kee, C. P. Classification techniques for predicting graduate employability. *International Journal on Advanced Science, Engineering and Information Technology, 8(4–2), 1712–1720. https://doi.org/10.18517/ijaseit.8.4-2.6832, 2018.*
- [12] Ministry of Higher Education and Scientific Research (MoHE) in Palestine (Arabic only), 2020. [online]. Available: <http://www.mohe.pna.ps>.
- [13] I. S. The International Standard Classification of Education (ISCED). In *Prospects (Vol. 5, Issue 2), 2012. [online]. Available: https://doi.org/10.1007/BF02207511.*
- [14] Machine learning in Python, 2020. [Online]. Available at <http://www.Scikit-learn.org>.
- [15] Valdez, A., Cortes G., Castaneda, S., and Laura, V. Development and Implementation of the Balanced Scorecard for a Higher Educational Institution using Business Intelligence Tools. (IJACSA) *International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017.*
- [16] Mahmud, Y., Shaeali, N. S., Mutalib, S., Comparison of Machine Learning Algorithms for Sentiment Classification on Fake News Detection. (IJACSA) *International Journal of Advanced Computer Science and Applications, Vol. 12, No. 10, 2021.*

# A Novel Stance based Sampling for Imbalanced Data

## A Case Study on COVID-19 Healthcare News

Isha Agarwal, Dipti Rana, Aemie Jariwala, Sahil Bondre  
Department of Computer Engineering, SVNIT, Surat, India

**Abstract**—While the world is suffering from coronavirus pandemic (COVID-19), a parallel battle with Infodemic, the proliferation of fake news online is also taking place. The spread of fake news during this global pandemic COVID-19 has dangerous consequences. This is the driving force behind this study. Relying on incorrect information obtained from the internet or social media can be fatal. According to a World Health Organization survey, at least 800 people have lost their lives because of COVID-19 misinformation during this time, highlighting the accurate automated classification of fake news. However, the data at disposal for classification is imbalanced. The Internet has a vast repository of authentic healthcare news, whereas Fake News on COVID-19 healthcare is not abundant. This imbalance leads to incorrect classification. The paper studies alternative approaches to text sampling. In this paper, we propose a stance based sampling method for balancing news data. The disparity between the title and content of news items is utilized to sample data points selectively and rectify the imbalance. The key findings are that the proposed stance-based sampling strategies enhance categorisation task performance consistently for varying degrees of imbalance. The proposed techniques can better detect misleading news in the health care sector.

**Keywords**—Fake news; healthcare; sampling; stance; COVID-19; imbalance

### I. INTRODUCTION

More than half of the global population now owns a smartphone, has internet access, and uses social media. There has been a 13.2 per cent rise in social media users by 2020. During the COVID19 outbreak, there was a tremendous spread of fake news and misinformation on a multitude of

health-related topics. The World Health Organization (WHO) coined the term "Infodemic" to characterize the spread of false information. This information apocalypse has deadly implications, which is why a system to identify misleading news is urgently needed. JS Brennen et al. identified the types of misinformation on COVID-19 [1].

Real news articles about health issues outweighed those that had been validated and labeled as fake, causing an imbalance in the news dataset. The most common solution to this problem is sampling to restore data balance. The two-class sampling problem for non-textual numeric data was explored and summarized by Japkowicz and Stephen in 2002 [2]. However, not much contribution has been made to textual data. This research uses stance to present a novel data sampling strategy for rebalancing the classes of news content in the healthcare sector (Fig. 1). In contrast to standard sampling strategies used to improve classification performance, the implications of stance-based classification for false news detection are examined.

The study begins by reviewing the necessary theoretical foundation and academic work in text preprocessing, feature extraction, stance identification, and textual sampling (Section II). Section III introduces a curated dataset for assessing the performance of the proposed algorithms. The training of a stance classifier, which is required for stance-based algorithms, is described in Section IV. The stance-based approaches are discussed in detail in Sections V and VI. The results of the algorithms are presented in Section VII, along with a comparative study of traditional approaches. Finally, a brief conclusion of the paper, along with the future scope of research, is laid out in the concluding section.

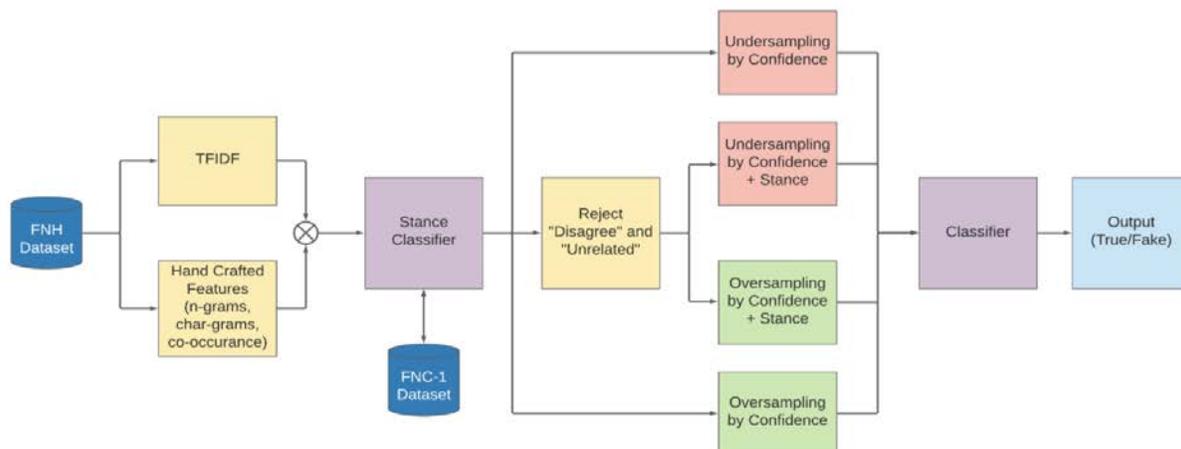


Fig. 1. Block Diagram for Sampling using Stance.

## II. BACKGROUND AND RELATED WORK

During the COVID-19 epidemic, various traditional and deep learning techniques for fake news detection are being studied. For training the textual data, important features need to be extracted, and thus, the textual data needs to be first preprocessed, followed by feature extraction.

### A. Textual Data Preprocessing

Within various studies and research, apart from tokenisation and stopword removal, authors have performed removal of HTTP URLs special characters [3][4][5]. In the study [6], the authors, in addition to the traditional preprocessing techniques, data augmentation using the back translation technique to increase the existing data is performed. The back-translation technique is the process in which the text is translated to its original language by converting it first into an intermediate language.

### B. Feature Extraction

Along with preprocessing, the main task involves feature extraction, after which the model is trained using traditional or deep learning classifiers. Within feature extraction, various methods have been used, to name the popularly used include TF-IDF, GLoVe, and Pre-trained BERT. For TF-IDF, different kinds of features are tested, including uni-gram, bi-gram, character level, etc. The studies [7][8] used these different TF-IDF representations at word-level, n-grams, etc., before feeding them to the classifier and obtained excellent results. Various studies [6][9][10][11][12] applied TF-IDF to convert the textual data into vector space and extract the important features. These studies showed a significant detection of fake news with an accuracy of 80-95%.

The limitation associated with TF-IDF is that it takes into account the occurrence of a particular word and not its grammatical meaning. This is where word-representation such as GLoVe and BERT shine. Stanford developed a global vector for word representation, termed GLoVe [13]. Each word is represented in a meaningful vector space where the cosine distance between two words depicts their similarity. In the studies [14][15], the authors applied an embedding layer using 300-dimensional pre-trained glove vectors. This layer could convert the tweet texts into a meaningful vector space. Dharawat et al. [11] utilised a 100-dimensional pre-trained glove vector along with various classifiers, and similarly, other studies [16][4] employ the same dimension vector for the feature extraction process.

Google developed a pre-training NLP technique, termed BERT [17]. It is based on an understanding of the context and relationship by learning text representation in both directions. There are two main models of BERT - BERT Base and BERT Large and mBERT is the BERT representation for multilingual representation. In the study [18], pre-trained BERT embeddings and mBERT have been utilized to extract features from tweets. Hossain et al. [19] have utilized pre-trained BERT embedding for understanding the similarity between misconceptions and tweets. Cheng et al. [20] used the BERT embedding for converting rumor texts into vector form. After BERT, the LSTM-based variational autoencoder [21] is

utilized to extract the important features. With this approach, a sufficient performance score was obtained. Various methods are utilized. However, these three embeddings are commonly used and help with providing efficient performance.

### C. Stance Detection

Stance detection is the process of identifying the stance (related, unrelated, etc.) from the textual data. It is identified through understanding the similarity of the headline and body of news content or article [22]. Common approaches involve training a labeled dataset with their stances, but a challenging task in this area includes stance detection without having the target values or no training data.

Lillie et al. investigated the topic of false news identification and stance classification and published their findings [23]. Echo chambers and model organism issues are two examples of difficulties that make collecting high-quality data challenging. Several methods for stance classification and fake news detection have been explored, but it has been difficult to compare their results because of different data and measures. One specific approach is very appropriate and interesting for the thesis project, which is the use of a Hidden Markov Model (HMM) in analysing rumours in microblog data, achieving very promising results. Augenstein et al. experiment with conditional LSTM encoding to build a representation of tweet dependent on target [24]. An additional change includes augmenting the conditional encoding along with bidirectional encoding for stance detection.

### D. Sampling Textual Data

Japkowicz et al. studied and unified all the previous approaches for solving the class imbalance problem using sampling and explained the nature of the problem by comparing the performance of the learning concept on parameters like complexity, training set size, and degree of imbalance [2]. A critical insight from the study was that class imbalance is not a problem because of the relative size of the small and large class, but it is only a problem when the size of its small class is too little for the complexity of the concept, i.e. when it contains minimal examples per subcluster. When each subcluster of the minority class contains many examples, accuracy remains high no matter the amount of imbalance or complexity of the concept. Textual data is a complex concept to learn, and the data distribution is sparse.

An active learning heuristic and representative sampling strategy is to read through the clustering structure of "uncertain" documents, reducing human effort in text classification tasks [25]. It also provides typical samples from which users can be polled to speed up SVM classifier convergence. This random sample includes more than one unlabelled document. Representative sampling was also compared to SVM active learning and random sampling by Zhao Xu et al. [25].

## III. DATASET

For training the model, a curated dataset for fake news in the healthcare domain is required. Within this paper, the FNH dataset has been used. It consists of the following features -

Title, Content, URL, and Publishing Date. This hand-curated dataset has been created using web scraping techniques from various fake/satire and true labeled websites. The statistics for the dataset are presented in the table (Table I), where true news instances supersede the fake news instances, thus creating a high imbalance in the news ecosystem.

TABLE I. FNH DATASET DISTRIBUTION

|       | Fake | True |
|-------|------|------|
| Count | 2424 | 7069 |

#### IV. TRAINING THE STANCE CLASSIFIER

For correcting the imbalance existing in the dataset, the stance approach has been chosen. Stance takes into account the textual similarity between the title and body content. Based on the similarity, we can gather its stance value and decide which instances of the particular class need to undergo sampling. This approach provides better insights on choosing instances to undergo sampling than the random traditional approach.

However, the FNH dataset has no stance labelled attribute. Introduction of the stance and its confidence for each instance of the dataset, a stance-labelled dataset is trained. In this paper, the FNC-1 dataset is used as the stance-labelled dataset. The training set is the entire FNC-1 dataset, and the testing set is the FNH dataset. A classifier works with the numerical data, and thus the textual data is represented in vector form.

##### Algorithm 1: Training Stance Classifier on FNC-1

**Input:** An annotated list of documents from FNC-1 dataset

**Output:** Classifier trained to identify stance between title and content of news article

**Begin:**

1. Create a vocabulary of words  $V$  from all text data
2.  $X := \{ \}$
3.  $N = \text{size}(\text{FNC-1})$
4. For every document  $d_i$  in FNC-1,  $i := [1,2,3,\dots,N]$ :
  - a.  $t_i := d_i.\text{title}$ ,  $c_i := d_i.\text{content}$
  - b.  $V_{t_i} := \text{TF-IDF}(V, t_i)$ ,  $V_{c_i} := \text{TF-IDF}(V, c_i)$
  - c.  $P_i := \text{co\_occurrence}(t_i, c_i)$
  - d.  $Q_i := \text{n\_grams}(t_i, c_i, n)$
  - e.  $R_i := \text{char\_grams}(t_i, c_i, n)$
  - f.  $S_i := \text{word\_overlap}(t_i, c_i)$
  - g.  $w_{t_i} := \text{word\_embedding}(t_i)$ ,  $w_{c_i} := \text{word\_embedding}(c_i)$
  - h.  $T_i := \text{cosine\_similarity}(w_{t_i}, w_{c_i})$
  - i.  $X_i := [V_{t_i}, V_{c_i}, P_i, Q_i, R_i, S_i, T_i]$
  - j.  $\text{append}(X, X_i)$
5.  $Y := \text{FNC1.stance\_labels}$
6. Create Instance of Multinomial Naive Bayes Classifier: MNB
7.  $\text{MNB.train}(X, y)$
8. Return MNB

**End**

Along with using TF-IDF (or word vectorization method), a hand-crafted vector space is created to emphasize the correlation between the headline and body of each document in a vectorized format. The hand-crafted vector space is a 28-dimensional vector space, and the distribution is explained in the table (Table II).

TABLE II. DISTRIBUTION OF VECTOR SPACE IN HAND-CRAFTED FEATURES

| Feature                                  | No. of vectors |
|------------------------------------------|----------------|
| Binary co-occurrence                     | 2              |
| Binary co-occurrence (stopwords removal) | 2              |
| N-grams                                  | 3              |
| Char-grams                               | 3              |
| Count-grams                              | 22             |
| Word-overlap                             | 1              |
| Word-embedding                           | 1              |

The TF-IDF vectors of the headline and the body, each 100 size vector are concatenated along with the handcrafted vectors is concatenated to give a 228 vector space for each document.

Within the FNC-1 dataset, there are five classes, and thus, Multinomial Naive Bayes takes into account of Bayes Theorem and provides the probability for the different classes for a single instance. Thus, the 228 vector space is subjected to a Multinomial Naive Bayes classifier to create the final trained model. The final trained model is then used for predicting the stance and confidence for the FNH dataset.

##### Algorithm 2: Generating Stance Values for FNH

**Input:** List of documents from FNH Dataset and an instance of MNB classifier from Algorithm 2

**Output:** Stance values for each document in the dataset

**Begin:**

1.  $X := \{ \}$
2.  $N := \text{size}(\text{FNH})$
3. For every document  $d_i$  in FNH,  $i := [1,2,3,\dots,N]$ :
  - a.  $t_i := d_i.\text{title}$ ,  $c_i := d_i.\text{content}$
  - b.  $V_{t_i} := \text{TF-IDF}(V, t_i)$ ,  $V_{c_i} := \text{TF-IDF}(V, c_i)$
  - c.  $P_i := \text{co\_occurrence}(t_i, c_i)$
  - d.  $Q_i := \text{n\_grams}(t_i, c_i, n)$
  - e.  $R_i := \text{char\_grams}(t_i, c_i, n)$
  - f.  $S_i := \text{word\_overlap}(t_i, c_i)$
  - g.  $w_{t_i} := \text{word\_embedding}(t_i)$ ,  $w_{c_i} := \text{word\_embedding}(c_i)$
  - h.  $T_i := \text{cosine\_similarity}(w_{t_i}, w_{c_i})$
  - i.  $X_i := [V_{t_i}, V_{c_i}, P_i, Q_i, R_i, S_i, T_i]$
  - j.  $\text{append}(X, X_i)$
4.  $\text{Stance\_Labels} := \text{MNB.predict}(X)$
5. return  $\text{Stance\_Labels}$

**End**

#### V. UNDERSAMPLING USING STANCE

For balancing the classes in undersampling, the instances of the majority classes are deleted till it is equal to the instances of minority classes. The deletion of the instances can be random, which is a traditional yet inefficient approach. Deleting the instances based on a systematic algorithm is an efficient approach.

In the previous section, the algorithm to acquire the stance label and confidence for each document is presented. The principle followed for undersampling is that the documents associated with low confidence should be subjected to deletion, which further resolves the imbalance.

In algorithm 3, the majority (true) class is sorted in descending order based on the confidence attribute, and the first N attributes are taken into consideration where N is equal to the number of instances belonging to the minority (fake) class.

---

**Algorithm 3: Undersampling based on Confidence**

---

**Input:** List of documents from FNH Dataset labeled with stance

**Output:** Undersampled data for classification task

**Begin**

1.  $D := \{FNH, Stance\_Labels\}$
2.  $True\_News := D.filter(d \text{ where } d.label = \text{"True"})$
3.  $Fake\_News := D.filter(d \text{ where } d.label = \text{"Fake"})$
4.  $True\_News := sort(True\_News, True\_News.stance.confidence, reverse=true)$
5.  $N := size(Fake\_News)$
6.  $Sampled\_True\_News := pick\_n(True\_News, N)$
7.  $Sampled\_Data := \{Sampled\_True\_News, Fake\_News\}$
8. return  $Sampled\_Data$

**End**

---

Along with the confidence attribute, the stance attribute has also been introduced within the FNH dataset. As the undersampling has been performed on the majority (true) class, the removal of stance attributes that are labeled as “disagree” or “unrelated” needs to be performed. The reason being that if a particular document is labeled true, then the headline and body needs to belong to the “agree” or “discuss” stance.

In Algorithm 4, after the deletion based on the stance values, sorting has been performed on the majority (true) class based on the confidence in the descending order. The first N instances are chosen where N is equal to the number of instances of the minority (fake) class.

---

**Algorithm 4: Undersampling based on Stance and Confidence**

---

**Input:** List of documents from Fake News Dataset labeled with stance

**Output:** Undersampled data for classification task

**Begin**

1.  $D := \{FNH, Stance\_Labels\}$
2.  $True\_News := FNH.filter(d \text{ where } d.label = \text{"True"} \text{ and } d.stance.label \neq (\text{"Disagree"} \text{ or } \text{"Unrelated"}))$
3.  $Fake\_News := FNH.filter(d \text{ where } d.label = \text{"Fake"})$
4.  $True\_News := sort(True\_News, True\_News.stance.confidence, reverse=true)$
5.  $N := size(Fake\_News)$
6.  $Sampled\_True\_News := pick\_n(True\_News, N)$
7.  $Sampled\_Data := \{Sampled\_True\_News, Fake\_News\}$
8. return  $Sampled\_Data$

**End**

---

The accuracy of the minority class is heavily weighted in evaluating performance since, in an unbalanced dataset, the minority class accuracy must improve. Thus, undersampling with stance variations is compared to the baseline, which was the original imbalance data, as well as the traditional undersampling approach.

From the graph (Fig. 2), it can be concluded that both approaches utilizing undersampling using stance supersede the traditional undersampling method and the baseline in performance as seen. Undersampling using stance and confidence performs the best as the imbalance ratio increases. This showcases that randomly choosing instances to undergo undersampling is an inefficient approach compared to utilizing the stance and confidence associated with each document.

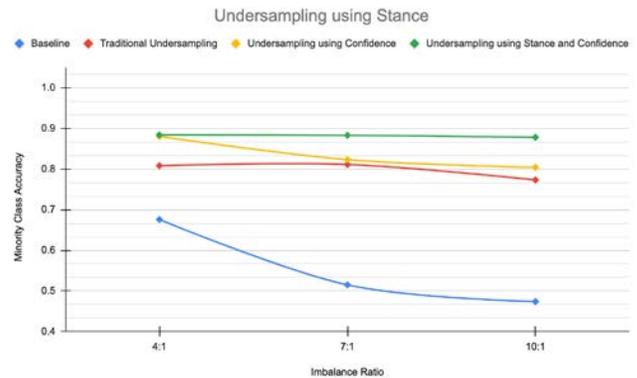


Fig. 2. Evaluation of Undersampling using Stance.

## VI. OVERSAMPLING USING STANCE

To balance the classes in oversampling, the minority class instances are oversampled until they are equal to the majority class instances. The oversampling method involves selecting a subset of minority class instances. These subsets are duplicated in a method that when these oversampled instances are added to the original minority instances, they equal instances of the majority class. The direct duplication could lead to overfitting, and hence it is important to choose an optimal number of subsets that undergo duplication to avoid overfitting.

Oversampling using stance uses the same base principle utilized in undersampling using stance. In Algorithm 5, first, sorting the minority (fake) class instances in descending order based on the confidence is performed.

The k integer which decides the subsets which will undergo oversampling is chosen in a way to avoid overfitting. In the case of the FNH dataset, the k chosen is 100 to keep the direct duplication of the subjects under 100. Choosing k within the range of [10, 50] requires the direct duplication of the subset to be done more than 150 times to resolve the imbalance. This leads to the overfitting of the data. However, choosing the k value to be greater than 100 will increase the time taken as larger subsets are chosen to undergo oversampling.

Once the top k instances are chosen from the minority class, they are subjected to oversampling such that the number of instances of both majority and minority classes is equal.

**Algorithm 5: Oversampling based on Confidence**

**Input:** List of documents from FNH Dataset labelled with stance and integer k

**Output:** Oversampled data for classification task

**Begin**

1.  $D := \{FNH, Stance\_Labels\}$
2.  $True\_News := D.filter(d \text{ where } d.label = "True")$
3.  $Fake\_News := D.filter(d \text{ where } d.label = "Fake")$
4.  $Fake\_News := sort(Fake\_News, True\_News.stance.confidence, reverse=true)$
5.  $Sampling\_Examples := pick\_n(Fake\_News, k)$
6.  $Sampled\_Fake\_News := Fake\_News + oversample(Sampling\_Examples)$
7.  $Sample\_data := \{True\_News, Sampled\_Fake\_News\}$
8. return  $Sample\_data$

**End**

For oversampling using stance and confidence, the same principle utilized for undersampling using stance and confidence has been used. After rejecting based on the stance value and sorting the minority class instances in the descending order based on their confidence, the first k subsets are chosen, which undergo oversampling (Algorithm 6). The method to choose the value of k has been explained in oversampling using stance.

**Algorithm 6: Oversampling based on Stance and Confidence**

**Input:** List of documents from FNH Dataset labeled with stance and integer k

**Output:** Oversampled data for classification task

**Begin**

1.  $D := \{FNH, Stance\_Labels\}$
2.  $True\_News := D.filter(d \text{ where } d.label = "True")$
3.  $True\_News := FNH.filter(d \text{ where } d.label = "Fake" \text{ and } d.stance.label != ("Disagree" \text{ or } "Unrelated"))$
4.  $Fake\_News := sort(Fake\_News, True\_News.stance.confidence, reverse=true)$
5.  $Sampling\_Examples := pick\_n(Fake\_News, k)$
6.  $Sampled\_Fake\_News := Fake\_News + oversample(Sampling\_Examples)$
7.  $Sample\_data := \{True\_News, Sampled\_Fake\_News\}$
8. return  $Sample\_data$

**End**

Oversampling with stance variations is compared to the baseline, which was the original imbalance data, as well as the traditional oversampling approach, and the priority is provided to the accuracy of the minority class. The reason is that within imbalanced data, the model tends to overfit, and the performance of the minority class is low.

Oversampling using stance supersedes the traditional oversampling method and the baseline in performance, as seen in Fig. 3. Oversampling using stance and confidence increases its performance and has a steady increase in the accuracy of the minority class across all imbalance ratios, while the traditional oversampling method scores reduce as it reaches a high imbalance ratio.

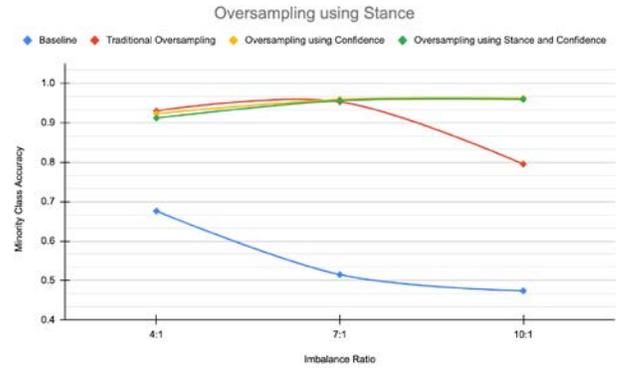


Fig. 3. Evaluation of Oversampling using Stance.

**VII. RESULTS**

For the evaluation purpose, the MCC score is taken into account. MCC is the only evaluation metric that considers all four quadrants of a confusion matrix, whereas Accuracy and Precision skew toward the positive class. To understand whether the model is overfitting by having high accuracy for the majority and low accuracy for the minority class or the model is balanced effectively, the accuracy of majority and minority class is both taken into account.

The evaluation metric is based on the confusion matrix and the MCC score. This has been done by averaging five trials for each method. The tables (Tables III to V) provide the performance of sampling using stances against the traditional sampling methods for the imbalance ratios 4:1, 7:1, 10:1.

For 4:1 imbalance ratio, it can be observed that oversampling methods supersede the undersampling methods in performance. Within the undersampling methods, the stance methods exceed in performance compared to the traditional method by a huge margin.

TABLE III. SIMULATION OF RESULT FOR RATIO 4:1

| Stance                             | Accuracy Majority Class | Accuracy Minority Class | MCC   |
|------------------------------------|-------------------------|-------------------------|-------|
| Baseline                           | 0.949                   | 0.677                   | 0.659 |
| Traditional Random Undersampling   | 0.884                   | 0.801                   | 0.685 |
| Traditional Random Oversampling    | 0.908                   | 0.932                   | 0.841 |
| Undersampling: Confidence          | 0.785                   | 0.876                   | 0.665 |
| Undersampling: Stance & Confidence | 0.872                   | 0.889                   | 0.76  |
| Oversampling: Confidence           | 0.921                   | 0.919                   | 0.84  |
| Oversampling: Stance & Confidence  | 0.925                   | 0.918                   | 0.843 |

Within the oversampling methods, the traditional oversampling method shows a minor improvement in performance compared to the oversampling using stance methods.

TABLE IV. SIMULATION OF RESULT FOR RATIO 7:1

| Stance                             | Accuracy Majority Class | Accuracy Minority Class | MCC   |
|------------------------------------|-------------------------|-------------------------|-------|
| Baseline                           | 0.978                   | 0.495                   | 0.569 |
| Traditional Random Undersampling   | 0.842                   | 0.814                   | 0.655 |
| Traditional Random Oversampling    | 0.935                   | 0.953                   | 0.889 |
| Undersampling: Confidence          | 0.835                   | 0.817                   | 0.653 |
| Undersampling: Stance & Confidence | 0.883                   | 0.894                   | 0.777 |
| Oversampling: Confidence           | 0.947                   | 0.959                   | 0.907 |
| Oversampling: Stance & Confidence  | 0.948                   | 0.955                   | 0.905 |

For 7:1 ratio, it follows the similar pattern as 4:1 ratio where oversampling methods supersede the undersampling methods in performance. Within undersampling methods, the undersampling using confidence showed a significant drop while undersampling using stance and confidence supersede in performance by a huge margin.

Within oversampling methods, the difference in performance for oversampling using stance and traditional method is very less. This showcases that with an increase in imbalance ratio, the oversampling using stance shows improvement in their performance.

For 10:1 ratio, both of the oversampling using stance variants supersedes in performance while the performance of traditional oversampling method reduces significantly. Within the undersampling methods, the undersampling using stance and confidence showcases steady improvement in performance with an increase in the imbalance ratio.

TABLE V. SIMULATION OF RESULT FOR RATIO 10:1

| Stance                             | Accuracy Majority Class | Accuracy Minority Class | MCC   |
|------------------------------------|-------------------------|-------------------------|-------|
| Baseline                           | 0.983                   | 0.466                   | 0.558 |
| Traditional Random Undersampling   | 0.833                   | 0.793                   | 0.625 |
| Traditional Random Oversampling    | 0.955                   | 0.863                   | 0.840 |
| Undersampling: Confidence          | 0.798                   | 0.818                   | 0.616 |
| Undersampling: Stance & Confidence | 0.872                   | 0.886                   | 0.758 |
| Oversampling: Confidence           | 0.949                   | 0.958                   | 0.908 |
| Oversampling: Stance & Confidence  | 0.953                   | 0.959                   | 0.914 |

## VIII. CONCLUSION AND FUTURE WORK

People are led to believe false facts about various health advice and medical treatments as a result of fake news. This creates a pressing need for accurate detection of fake news in healthcare. The proposed framework focuses on improving the performance of fake news detection in order to address these issues. Because the number of true articles in this work outnumbers the number of fake articles, stance has been used for the text sampling method, both undersampling and oversampling.

Understanding the relationship between the headline and the article's content is essential in stance classification. The FNH dataset has been trained to obtain their respective stance labels and confidence. Two approaches have been proposed. Stance-based undersampling and stance-based oversampling were carried out using these variations. These proposed approaches demonstrated a significant improvement in overall detection performance when implemented with various imbalance ratios compared to traditional methods.

Apart from increasing the performance using stance by resolving balance, the broader implications of the paper also highlight the unique method of converting the textual data into vector space highlighting the similarity between the title and body content of the document which further was utilized grabbing the stance and confidence attribute for each document.

Future work can be extended by training different classifiers for stance detection. Experiments can also be further carried out considering the tuning of configuration parameters for the rate of sampling, etc.

### REFERENCES

- [1] Brennen JS, Simon F, Howard PN, Nielsen RK. Types, sources, and claims of COVID-19 misinformation. Reuters Institute. 2020 Apr 7;7(3):1.
- [2] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. *Intelligent data analysis*. 2002 Jan 1;6(5):429-49.
- [3] R.Kaliyar,A.Goswami,P.Narang, "A Hybrid Model for Effective Fake News Detection with a Novel COVID-19 Dataset".
- [4] Wani, I. Joshi, S. Khandve, V. Wagh, R. Joshi, "Evaluating Deep Learning Approaches for Covid19 Fake News Detection".
- [5] Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, W. Ren, Y. Zhou , "Transformer-based Language Model Fine-tuning Methods for COVID-19 Fake News Detection".
- [6] J.Ayoub,X.Yang,F.Zhou, "Combat COVID-19 Infodemic Using Explainable Natural Language Processing Models".
- [7] M. K. Elhadad, K. F. Li, F. Gebali, "Detecting Misleading Information on COVID-19".
- [8] Mahlous, A. Al-Laith , "Fake News Detection in Arabic Tweets during the COVID-19 Pandemic".
- [9] A.Koirala , "COVID-19 Fake News Classification using Deep Learning".
- [10] P.Patwa,S.Sharma, S.PYKL,V.Guptha, G.Kumari, M.S.Akhtar,A.Ekbal, A.Das,T.Chakraborty, "Fighting an Infodemic: COVID-19 Fake News Dataset".
- [11] A.Dharawat,I.Lourentzou,A.Morales,C.Zhai, "Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation".
- [12] L. Alsudias, P. Rayson , "COVID-19 and Arabic Twitter: How can Arab World Governments and Public Health Organizations Learn from Social Media?"

- [13] J.Pennington, R.Socher, C.Manning , “GloVe: Global Vectors for Word Representation”.
- [14] M.Elhadad, K.Li, F.Gebali, “An Ensemble Deep Learning Technique to Detect COVID-19 Misleading Information”.
- [15] S.Kumar, K.M.Carley, “A Fine-Grained Analysis of Misinformation in COVID-19 Tweets”.
- [16] L.Cui,D.Lee , “CoAID: COVID-19 Healthcare Misinformation Dataset”.
- [17] J.Devlin, M.W.Chang, K.Lee, K.Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”.
- [18] Kar, M. Bhardwaj, S. Samanta, A. P. Azad, “No Rumours Please! A Multi-Indic-Lingual Approach for COVID Fake-Tweet Detection”.
- [19] T.Hossain, “COVIDLIES: Detecting COVID-19 Misinformation on Social Media”.
- [20] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, P. Bogdan, “A COVID-19 Rumor Dataset”.
- [21] M. Cheng, S. Nazarian, P. Bogdan, “VRoC: Variational Autoencoder-aided Multi-task Rumor Classifier Based on Text”.
- [22] Rajendran, B. Chitturi, P. Poornachandran, “Stance-In-Depth Deep Neural Approach to Stance Classification”.
- [23] Anders Edelbo Lillie and Emil Refsgaard Middelboe. “Fake news detection using stance classification: A survey”. In:arXiv preprint arXiv:1907.00181(2019).
- [24] Augenstein, T. Rocktäschel, A. Vlachos, K. Bontcheva , “Stance Detection with Bidirectional Conditional Encoding”.
- [25] Zhao Xu et al. “Representative sampling for text classification using support vector machines”. In:European conference on information retrieval. Springer. 2003, pp. 393–407.

# Energy Efficient and Quality-of-Service Aware Routing using Underwater Wireless Sensor Networks

P. Sathya<sup>1</sup>, P. Sengottuvelan<sup>2</sup>

Ph.D Research Scholar<sup>1</sup>, Associate Professor<sup>2</sup>,

Department of Computer Science, Periyar University PG Extension Centre, Dharmapuri, Tamilnadu, India<sup>1,2</sup>

**Abstract**—In current years, there has been an increasing attention in Underwater wireless sensor networks (UWSNs). Underwater sensor networks (USNs) can be applied for many various purposes. To address the routing issue, the Cuckoo Search Optimization Algorithm with Energy Efficient and QoS Aware (CSOA-EQ) based routing methods have been proposed in this chapter. Every application is important in its own right, but some of them can help improve sea investigation to meet a variety of underwater applications, such as a catastrophic event alert system (such as torrent and seismic monitoring), supported navigation, oceanographic data collection, and underwater surveillance, ecological applications (such as the nature of organic water and contamination monitoring), modern applications (such as marine investigation), and so on. For example, sensors can assess specific metrics, such as base intensity and securing pressure, to monitor the auxiliary nature of the securing environment in offshore engineering applications. UASNs have also improved our understanding of underwater environments, such as climate change, underwater creature life, and the number of inhabitants in coral reefs.

**Keywords**—Underwater wireless sensor networks (UWSNs); QoS aware (CSOA-EQ); underwater environments

## I. INTRODUCTION

Sensor Networks have emerged as a potential examination topic this year. In these types of networks, the routing issue is a critical component that must be handled in order to extend the life of the organisation. Because of the number of sensor nodes in the organisation, routing becomes increasingly unpredictable as the size of the organisation grows [1]. Sensor nodes in Wireless Sensor Networks are highly reliant on memory, processing power, and battery life.

Surface buoys operational with GPS can acquire their regions, as seen in Fig. 1. The numeral of beacon nodes be far lower than the amount of obscure nodes. Because beacon nodes have more energy and a communication range of roughly 200 metres, they can legitimately communicate with buoys and have more neighbours. Furthermore, compared to hidden nodes, beacon nodes have more equipment assets and superior figuring capacity, allowing them to do more. The obscure nodes are slightly less expensive, and they aren't required to waste energy [2]. An obscure hub's communication radius is roughly 100 metres, and it is unable to officially communicate with the buoys. It can normally only associate with its immediate (usually one-jump) neighbors [3]. The obscure nodes can achieve nearby positioning to successfully participate in the organization exercises thanks to local data exchange among themselves and nearby beacon nodes.

Optimization is everywhere, so there are many applications for this paradigm [4]. In practically all technical and industrial applications, we are continually looking for ways to improve anything, whether it is to minimize cost and energy consumption or to boost benefit, yield, execution, and effectiveness [5]. Because resources, time, and money are all limited in practice, optimization is undoubtedly more important. Because most real-world applications have clearly more complex variables and parameters influencing how the framework functions, making the greatest use of any given assets needs a paradigm shift in logical reasoning. The essential components of the optimization cycle for each optimization challenge are the optimization algorithm, an effective numerical test system, and a realistic-portrayal of the physical cycles we want to illustrate and optimize [6]. After we have a good model, the general computation costs are dictated by the optimization strategies used for search and the numerical solver utilized for simulation. To address the routing issue, Cuckoo Search Optimization Algorithm with Energy Efficient and QoS Aware (CSOA-EQ) based routing methods have been proposed [7]. A wireless sensor network (WSN) is made up of self-contained sensor nodes. These sensors are extremely small. They are widely disseminated in large numbers [8]. These sensor nodes are intelligent and successful, providing a very wonderful and adaptive network where regular wired and wireless networks are unable to deliver. WSN is used in a variety of engineering applications, including monitoring unfenced far margins, terrorist development in high-altitude backwoods territories, and LPG pipe lines put in deep water. The sensor node detects or monitors movement or functions in a network region or area and delivers information to the base station.

A QoS aware routing provides optimality when sensor nodes are transmitted in underwater sensing applications, such as monitoring LPG pipelines [9]. As a result, this research also focuses on QoS-aware routing. The compromise between network QoS assurance and network lifetime is a fundamental issue in QoS aware routing, i.e., the presentation of QoS aware routing and the implementation of Energy efficient routing are diametrically opposed. The main goal is to maintain the energy-efficient and QoS-aware routing running smoothly. This chapter describes WSN execution and compares it to ACO and PSO's suggested CSOA-EQ execution in terms of normal number of ways, energy consumption, and normal parcel delay [10]. The proposed CSOA-execution EQ's is considerably enhanced when compared to ACO and PSO.

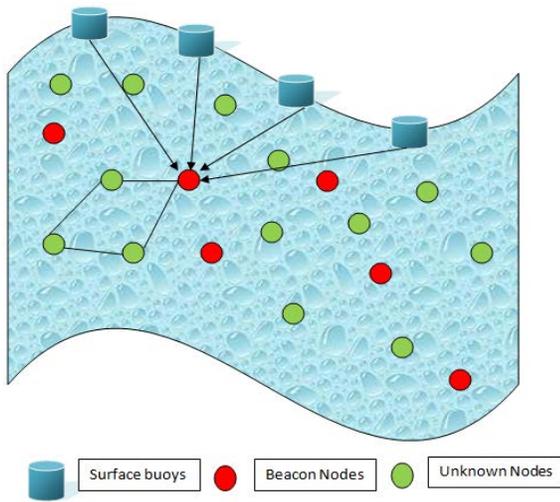


Fig. 1. Schematic Diagram of UWSN.

## II. LITERATURE REVIEW

Sudip Misra, Anudipa Mondal, and Ayan Mondal (2019) introduced DATUM, a game theory-based dynamic topology control approach for enhancing throughput and network lifetime in UWMSNs within the sight of interactive media sensor nodes with the least network latency [1]. This author uses a cooperation game theoretic approach to select the best combination of methods for limiting postponement and amplifying throughput, as well as the best transmission power for extending network lifetime. Luis M. Pessoa, Cândido Duarte, Henrique M. Salgado, Vasco Correia, Bruno Ferreira, Nuno A. Cruz, and Anibal Matos (2019) evaluate the long-term deployment feasibility of a large-scale network of abandoned underwater sensors, with power provided by autonomous underwater vehicles (AUVs) in periodic visits [2]. They conduct a versatile analysis to determine the size of network that can be supported by a single AUV, both in terms of total number of sensors and partition separation between sensors.

## III. EXISTING METHODOLOGIES

### A. Ant Colony Optimization Algorithm

The field of "Ant Algorithm" investigates how models derived from observations of real ants' behavior stimulate the development of novel algorithms for solving distributed control problems via optimization [11]. The basic idea is that self-organizing standards, which take into account the profoundly co-ordinate behavior of actual ants, may be utilized to co-ordinate populations of artificial agents working together to solve computing problems. Various forms of ant algorithms, including as seeking, division of labour, brood arranging, and co-employable vehicle, have been enlivened by a few distinct aspects of ant province behavior. "One of the better examples of ant algorithms is "Ant Colony Optimization (ACO)"[12]. Because the forward ants in the ACO strategy are sent to no specific destination node in the essential algorithm, sensor nodes must communicate with one another and every node's routing tables must contain the IDs of all sensor nodes in the neighbourhood as well as the journalist levels of the pheromone trail, making the ACO

strategy more energy efficient. In big networks, this can be an issue because nodes would require a lot of memory to save all of the data about the neighbourhood.

### B. Particle Swarm Optimization

Molecule Swarm Optimization (PSO) was developed by Kennedy and Eberhart in 1995, based on a range of natural behaviours such as fish and flying creature learning. PSO has since sparked a plethora of new interests and structures an exhilarating, ever-expanding study topic known as swarm insight. This method searches the spaces of a target work by changing the trajectories of individual agents, termed particles, in a quasi-stochastic manner as piecewise ways determined by positional vectors [13]. Particle swarm optimization (PSO) mathematical expression is followed[14]. Assuming a Dimensional search space, m particles form a group. The position of the particle I in the search space is  $x_i$  and Vector is  $x_i = (x_{i1}, x_{i2}, L, x_{iD})^T$  and the flying speed is  $V_i = (v_{i1}, v_{i2}, L, v_{iD})^T$ . The individual extremum of particle I is  $P_i = (P_{g1}, P_{g2}, L, P_{gD})^T$ . In the formula g is the number of fitness optimum in the group [13]. The particles are iterative operation according to the following formula (3) and (4) and schematic diagram is shown in Fig. 4.

$$v_{id}^{k+1} = \omega v_{id}^k + c_1 \cdot r_1 \cdot (p_{id}^k - x_{id}^k) + c_2 \cdot r_2 \cdot (g_{id}^k - x_{id}^k)$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^{k+1}$$

$\omega$  is non-unconstructive number, called the inertia weight (intertia weight), Its role is not to adjust the algorithm of global and local search ability of balance.  $I=1, 2, \dots, m$   $d=1, 2, \dots, D$ . Acceleration constant  $c_1$  and  $c_2$  (acceleration constant) is a non negative;  $r_1$  and  $r_2$  is a random number between.  $v_{id} \in [-v_{max}, v_{max}]$ ;  $v_{max}$  is a constant, setting by the user;  $k=1, 2, \dots$  is the number of iterations [15].

## IV. PROPOSED METHOD

### A. Network Architecture

First, we'll go through the network architecture employed in this chapter. The sensor network in this scenario is a Big Network (BN), with the water's surface as the top surface and the seabed as the bottom surface. This underwater sensor network is organized into miniature crow nests, as seen in Fig. 2 (CNs). Fig. 1 shows the network architecture of the suggested model.

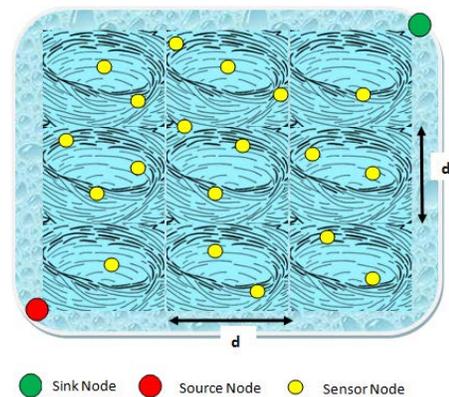


Fig. 2. Network Architecture of the Proposed Model.

A source node should be at the lower left corner of the BN, whereas a sink node should be in the upper right corner. Allow  $N_a$  sensor nodes equipped with audio modems to communicate with sonobuoys positioned at arbitrary depths. These nodes are all stagnant and do not have any water flowing through them. Each node in the network has the same main energy, transmission power, and range, as well as symmetric interactions between nodes. The sink node is assumed to also be equipped with an offshore level sonobuoy, and that each node is aware of both its own and the sink node's location. Furthermore, as seen in Fig. 2, there are several nodes within a CN. Furthermore, as seen in Fig. 2, there are many nodes within a CN that store the CN that they have a place with. The length of the CN is commonly referred to as  $d$ , and it is entirely dependent on  $r$ , which is the sensor node exchange radius. Similarly, the duty cycle technique assumes that each nodes condition varies independently of the state of the others. Because each node is awake for a short period of time before going to sleep for the remainder of the time, this is the simplest type of obligation cycle. There is no need for global synchronization because each sensor node can maintain track of its own and the sink node's locations.

### B. CSOA-EQ Algorithm

Similarly, because the routing database does not exist, no RAM is utilized to save the path. Additionally, the multiple-routing method will be combined with geo-routing to improve the reliability of packets being received efficiently, and packets will be dispatched from many routes at the same time. Furthermore, the duty cycle instrument allows the nodes to rest occasionally to conserve energy while no data is being transmitted. The network is said to be a Big Network separated into little crow homes. In the proposed technique; each sensor node can communicate directly with all of its vertex, edge, and surface adjacent nodes. A CN within a BN, as seen in Fig. 3, node  $u$ , is an outstanding illustration of this type of CN.

In Fig. 4, a flowchart of the proposed method is exposed.

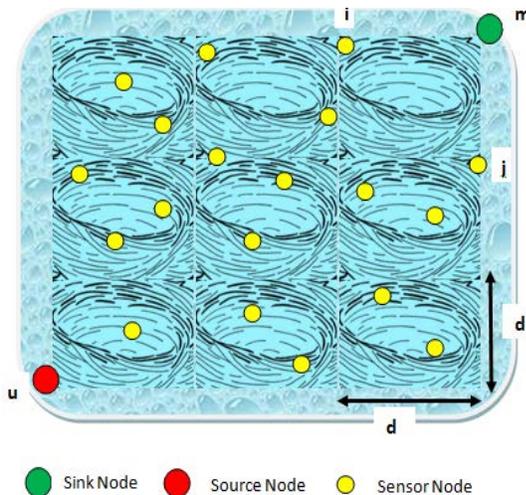


Fig. 3. Relation between  $r$  and  $d$  in the Network.

### C. Proposed CSOA-EQ Algorithm

Similarly, the routing database does not exist, so no memory is used in the process of storing the path. Furthermore, the multiple-routing process will be used with geo-routing to increase the reliability of packets being received effectively, and packets will be dispatched off the destination at the same time from numerous routes. In addition, the duty cycle instrument allows the nodes to occasionally rest in order to save energy while no data is being delivered. It's thought that the network is a Big Network divided into small crow homes. To prevent packet flooding, each node's diffusion radius is defined, and every node can only transfer data within this radius. Furthermore, in the suggested algorithm, node selection is settled in a later step depending on the node's reasonableness.

Every sensor node in the proposed algorithm can communicate directly with all of its vertex-adjacent, edge-adjacent, and surface-adjacent nodes. It should be noted that a CN within the BN, as shown in Fig. 2, is an excellent example of such a CN. Similarly, node  $m$  is the farthest node with whom  $u$  can have a direct conversation. As a result, the Base Exchange radius of node  $u$  is determined by the Euclidean distance between  $u$  and  $m$ .

As a result, the following equation determines the relationship between the length of the CN edge ( $d$ ) and the transfer radius of the sensor nodes ( $r$ )

$$(2d)^2 + (2d)^2 + (2d)^2 = r^2 d = \frac{r}{\sqrt{12}} \quad (1)$$

- First, as shown in Fig. 1,  $N_a$  nodes are haphazardly spread at various depths of  $H$  in BN space to depict the suggested technique. Based on the measure of  $d$ , the BN is then partitioned into various CNs. Using range-based or without range localization techniques; each node should also know its own location and the location of the sink node. Fig. 5 illustrates this. Furthermore, each node must be aware of which CN it belongs to. The fact that each of these nodes has the same fundamental energy is crucial. The routing cycle begins in two steps after the fundamental design of the nodes:
- Stage 1: The initial phase of the routing for the source node  $u$  is discovering all of the node's nearby Crow Nest (CN) subunits. A CN can have vertex-adjacent, edge-adjacent, and surface-adjacent neighbours, as previously stated. We select CNs that are closer to the sink than the current Crow Nest (CN) and have at least one awake node from all nearby CNs. This is to prevent the course from being bypassed in vain and the sink way from being escaped. The nodes in these CNs are then recorded and assessed as possibilities for the next stage. Following the shaping of a large number of applicant nodes, the second routing stage begins with the CSOA-EQ algorithm, as shown below.

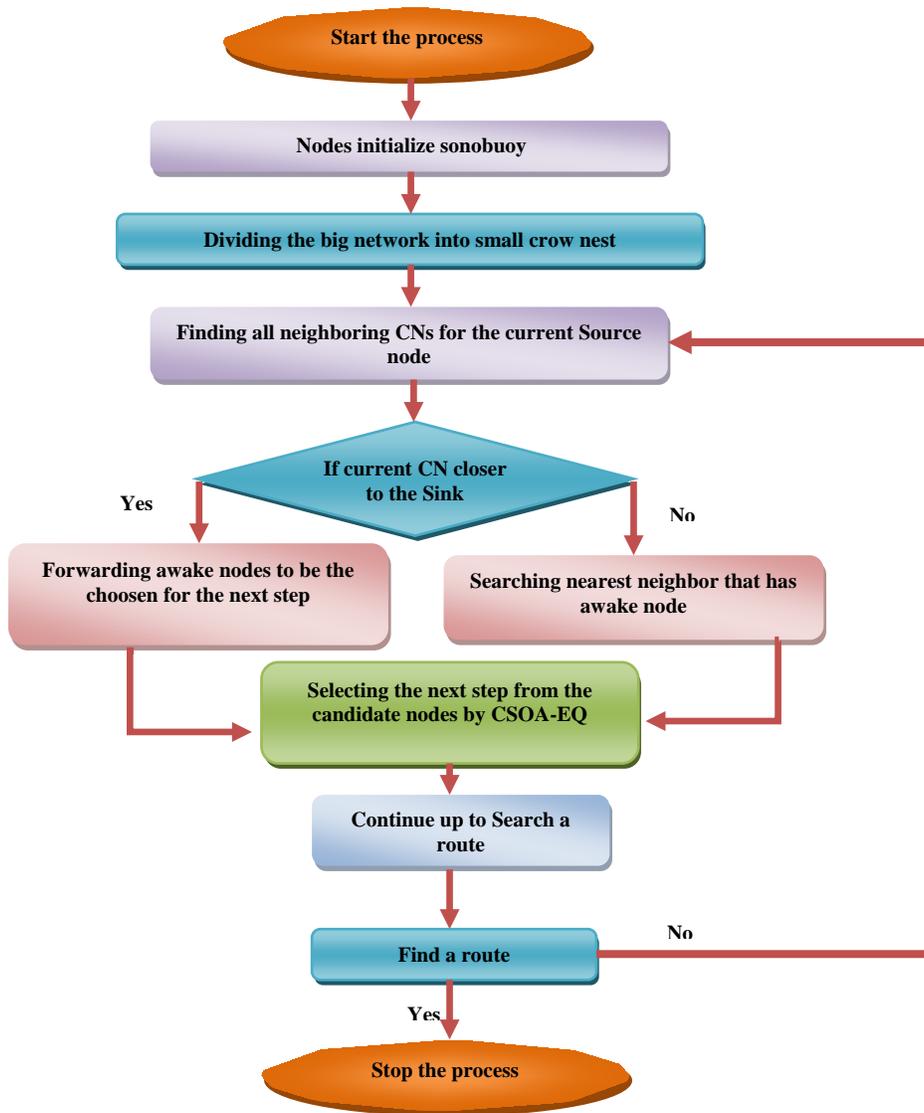


Fig. 4. Flowchart of the Proposed Method.

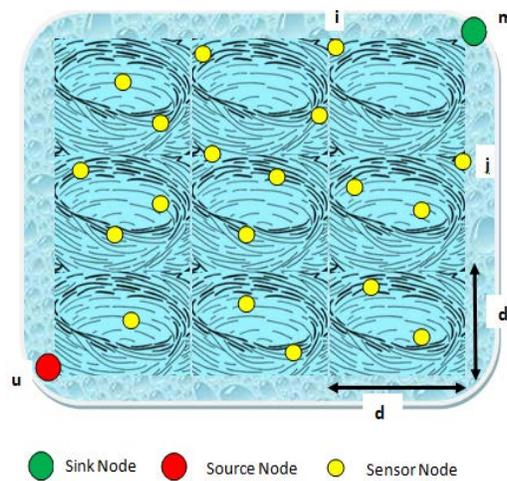


Fig. 5. Relation between  $r$  and  $d$  in the Network.

- Stage 2: First, mother cuckoos are horizontally distributed across the search field using the Cuckoo Search Optimization Algorithm. The following advancements are then carried out in order to achieve the optimal precision in optimization.
- Step 1: A random number of eggs is allotted to each mother cuckoo. The equation then determines the Egg Laying Radius (ELR) of each bird in relation to this number.

$$ELR = \beta \times \frac{\text{Number of current cuckoo's eggs}}{\text{Total number of eggs}} \times (\text{var}_{hi} - \text{var}_{low}) \quad (2)$$

Where  $\beta$  is a positive integer that represents the maximum ELR value. It's that each bird can only lay eggs within a certain radius.

- Step 2: Mother cuckoos lay their eggs in this area. After the laying process, the profit value of all mother cuckoos and egg is calculated using the cost function (described below). Then, in order to keep the cuckoo population under control, we destroy eggs with a lower yield. The laying process is completed at this point, and the eggs are grown and matured over time.
- Stage 2: This is where mother cuckoos lay their eggs. The benefit estimation of all mother cuckoos and eggs is calculated after the laying cycle using the cost work given below. To regulate the most extreme population of cuckoos, we kill eggs with a lower yield at that time. Near the end of this stage, the laying cycle is completed and the eggs are formed and developed.
- Stage 3: The presence of young cuckoos creates a new population of cuckoos. The algorithm's current mothers are young cuckoos. The cuckoos' dwelling areas are then grouped using the K-mean grouping technique. Because the search field and number of optimization are not large in this chapter, K is assumed to be 1.
- Stage 4: Following bunching, the benefit of each group is calculated, and the bunch with the highest benefit is presented as the best place for cuckoo relocation. In the same way, the cuckoo with the highest profit in this group is chosen as the worldwide ideal.
- Stage 5: If the algorithm arrives at the ideal combination, it stops; in any case, these means are reshaped. The issue in the CSOA-EQ algorithm is the closeness of the cuckoo's populace to one another. In the event that the cuckoo populace is near the greatness of the intermingling coefficient indicated in the algorithm, the algorithm has arrived at its assembly.

It ought to be noticed that CSOA-EQ yield is the global optimal arrangement which is accomplished by adjusting it to the neighboring node with the most elevated net revenue. The above advances proceed until there are no different nodes or CNs that gives routing conditions.

The following bounce node is best picked by the CSOA-EQ algorithm in this chapter based on the node with the maximum energy() in the objective CN. The distance between the current node and the chosen node, the method misfortune

based on the propagation delay, and the measure of the current node's residual energy are all factors that influence this decision. Regardless of the standards, the energy required to send, rest of the node's energy when data is shipped off this node, and the node's underlying energy are all taken into account in this cost calculation. Similarly, because exponential-sine capabilities are ones in which minor changes in parameters can result in dramatic changes in the capacity's implications, the cost work for selecting the next bounce node is as follows:

$$\dot{v}_{ij} = \frac{\mu}{D} + \frac{\kappa}{L} + \tau C_{ij} + \zeta E_{resi} \quad (3)$$

In Eq. (3),  $\mu$ ,  $\kappa$ ,  $\tau$ ,  $\zeta$ , they're constant coefficients with a sum of 1 that's utilized to manage the weight in the cost function.  $C_{ij}$  is also defined by.

$$C = E_{ij}^{rem} \exp\left[1/\sin\left(\pi - \frac{\pi E_{ij}^{rem}}{E_0}\right)\right] \quad (4)$$

In the above relation  $E_{ij}^{rem}$  is the remaining energy of node I if sent to node j,  $E_{ij}^{rem}$  is the current remaining energy of node i and  $E_0$  is the initial energy of node i. Accordingly, based on the relation, the optimal selection of the next step for node I through CSOA-EQ algorithm based on the node with the highest energy in the crow nest is in the crow nest is made by.

$$J = \text{Arg max}_{j \in \{S, N_d\}} (\omega_{ij}) \quad (5)$$

Here,  $N_d = \alpha N_a$  is a set of all awake nodes in the network according to the duty-cycle mechanism, Where nodes and  $0 < \alpha \leq 1$  is the duty-cycle parameter. It is worth nothing that the larger  $\alpha$  becomes, the more paths there are for the next selection, and so the network can route information from the source to the sink through different paths.

#### D. Time Complexity

As shown in Algorithm 2, CSOA-EQ for temporal complexity is made up of five key phases. Here's a quick rundown of each stage and its associated time complexity:

- Step 1: The first stage, according to Equation, is to calculate the dependability constraint score (6). The temporal complexity of this step can be estimated as follows:  $O(|T||AC|)$ .

$$\sum_{\forall C_j \in A_C} Y_{i,j} X_j \geq K; \forall v_i \in T \quad (6)$$

- Step 2: The second step is to calculate the time constraint score using Equation (7). This step takes a long time:  $O(|T||AC|)$ .

$$\sum_{\forall C_j \in T} Y_{i,j} X_j \frac{\omega_i}{\sum_{\forall C_j \in A_C} Y_{i,t} X_t} \leq \frac{w}{(k-1)}; \forall C_j \in A_C \quad (7)$$

- Step 3: The third step is to calculate the maximum distance between sensors and controllers using Equation (8). the time complexity of this step is:  $O(|T||AC|)$ .

$$L_{v_i}^* = \max_{\forall C_j \in A_C} \{Y_{i,j} X_j l^*(v_i, C_j)\} \quad (8)$$

- Step 4: The fourth phase entails sorting all of the eggs into fitness categories. The temporal complexity of this phase can be represented as  $O(\text{no of eggs}) \times N_{\text{pop}} \log N_{\text{pop}}$ .
- Step 5: The fifth stage entails sorting mature cuckoos based on their fitness values, which is a lengthy process:  $O(N_{\text{pop}} \log N_{\text{pop}})$ .

Hence, the time complexity can be written as  $O(N_{\text{pop}} \log N_{\text{pop}} + |T||AC|)$ . Furthermore, while increasing network size does not result in a significant change in  $N_{\text{pop}}$ , it can be concluded that  $|T||AC|$  has upper hand on  $N_{\text{pop}} \log N_{\text{pop}}$ . As a result, Cuckoo-overall PC's time complexity is:  $O(|T||AC|)$ .

### V. EXPERIMENTAL RESULTS

#### A. Number of Sensor Nodes on Average Number of Paths Found

Table I compares the number of sensor nodes to the average number of pathways discovered, which explains the ACO, PSO, and CSOA-EQ results. When the existing approaches and the new CSOA-EQ are compared, the proposed method produces better results (Fig. 6).

TABLE I. NUMBER OF SENSOR NODES ON AVERAGE NUMBER OF PATHS FOUND

| Number of sensor nodes | ACO   | PSO   | Proposed CSOA-EQ |
|------------------------|-------|-------|------------------|
| 200                    | 1.255 | 1.468 | 1.861            |
| 400                    | 2.216 | 2.365 | 2.632            |
| 600                    | 2.945 | 3.014 | 3.265            |
| 800                    | 3.465 | 3.984 | 4.147            |
| 1000                   | 4.875 | 4.971 | 5.852            |

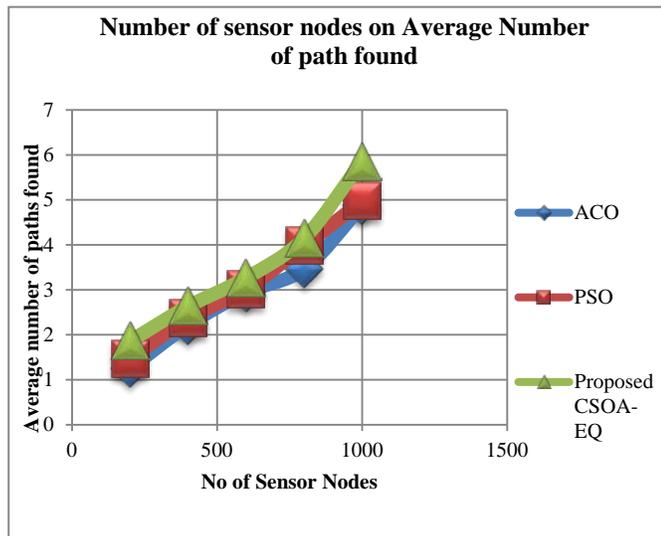


Fig. 6. Sensor Nodes on Average Number of Paths Found.

The evaluation table of Number of sensor nodes on average number of paths detected explains the ACO, PSO, and recommended CSOA-EQ values. When the results of existing approaches and the proposed method are compared, the

suggested CSOA-EQ method comes out on top. We can also observe that as the number of nodes grows, so does the number of paths detected, resulting in enhanced network dependability and speed in identifying the path within the network.

#### B. Number of Sensor Nodes on Energy Consumption

Table II shows a comparison table of the number of sensor nodes on the average number of pathways identified, which explains the ACO, PSO, and CSOA-EQ results. When comparing existing approaches to the proposed CSOA-EQ, Fig. 7 shows that the proposed CSOA-EQ uses less energy.

The comparison table of number of sensor nodes on average number of paths detected explains the ACO, PSO, and proposed CSOA-EQ values. When comparing the existing and suggested approaches, the proposed CSOA uses the least amount of energy.

#### C. Average Packet Delay

Table III provides the Average Packet Delay Comparison Table, which illustrates the differences in ACO, PSO, and CSOA-EQ values. Fig. 8 shows that when existing approaches are compared to the proposed CSOA-EQ, the suggested method produces better results.

The ACO, PSO, and suggested CSOA-EQ values are explained in the Comparison chart of Number of sensor nodes on average packet latency. When comparing the outcomes of the existing approaches and the proposed method, the suggested CSOA-EQ provides better results.

TABLE II. NUMBER OF SENSOR NODES ON ENERGY CONSUMPTION

| Number of sensor nodes | ACO   | PSO   | Proposed CSOA-EQ |
|------------------------|-------|-------|------------------|
| 200                    | 9.024 | 8.802 | 6.789            |
| 400                    | 8.632 | 7.744 | 5.065            |
| 600                    | 8.125 | 7.062 | 4.732            |
| 800                    | 7.851 | 6.745 | 4.487            |
| 1000                   | 6.974 | 6.196 | 2.954            |

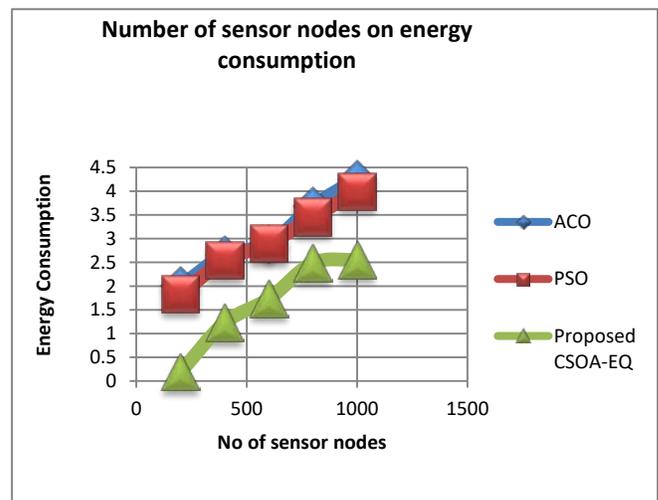


Fig. 7. Sensor Nodes on Average Number of Paths Found.

TABLE III. AVERAGE PACKET DELAY

| Number of sensor nodes | ACO   | PSO   | Proposed CSOA-EQ |
|------------------------|-------|-------|------------------|
| 200                    | 1.999 | 1.821 | 0.189            |
| 400                    | 2.654 | 2.544 | 1.225            |
| 600                    | 2.891 | 2.892 | 1.732            |
| 800                    | 3.684 | 3.456 | 2.487            |
| 1000                   | 4.258 | 3.996 | 2.454            |

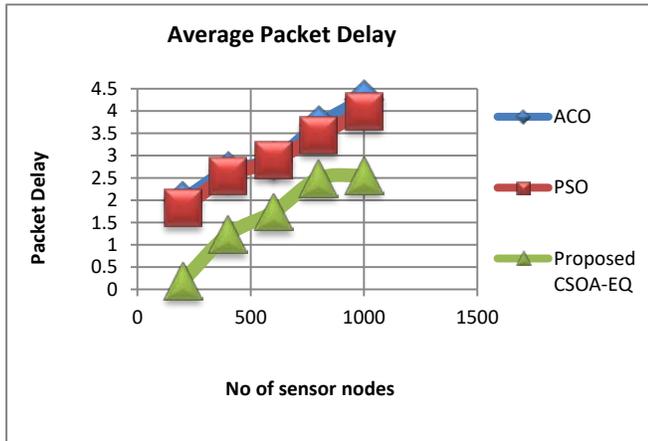


Fig. 8. Sensor Nodes on Average Packet Delay.

## VI. CONCLUSION

Cuckoo Search Optimization Algorithm with Energy Efficient and QOS Aware (CSOA-EQ) based routing methods have been proposed in this research to address the routing issue while attempting to handle these constraints. The proposed approach is also utilized to boost the likelihood of items being delivered successfully. By using duty cycle approach, nodes are also placed to sleep on a regular basis to save energy while no data is delivered. By introducing a differential equation in the CSOA-EQ algorithm, path selecting in the proposed technique is performed based on energy utilization and residual energy of the current node. When designing a network path for a WSN, fuel efficiency is a crucial consideration. The effectiveness of CSOA-EQ is appropriate in many application areas, as evidenced by its use in WSN routing issues. In this paper, we present a routing technique just on Cuckoo Search Optimization Algorithm with Energy Efficient and QOS Aware (CSOA-EQ), and also a probability of route selection based on pheromone and residual energy. We established that the proposed strategy is more energy efficient through simulated results. As part of future research, the CSOA-EQ algorithm will be used to increase network longevity, energy usage, and average packet latency.

## REFERENCES

[1] SudipMisra, AnudipaMondal, and AyanMondal (2019), "DATUM: Dynamic Topology Control for Underwater Wireless Multimedia Sensor Networks", Electronic ISSN: 1558-2612, DOI: 10.1109/WCNC.2019.8885632, IEEE.

[2] Luís M. Pessoa, Cândido Duarte, Henrique M. Salgado, Vasco Correia, Bruno Ferreira, Nuno A. Cruz and Anibal Matos (2019), "Design of an underwater sensor network perpetually powered from AUVs", Electronic ISBN: 978-1-7281-1450-7, DOI: 10.1109/OCEANSE.2019.8867273, IEEE.

[3] Sai Wang and Yoan Shin (2019), "3D-Deployment of Magnetic Induction Relays in Underwater Sensor Networks", DOI: 10.1109/ICOIN.2019.8718105, Electronic ISBN: 978-1-5386-8350-7, IEEE, pp.222-226.

[4] Zhenghao Xi, XiuKan, Le Cao, Huaping Liu, GunasekaranManogaran, George Mastorakis, Constandinos and X. Mavromoustakis (2019), "Research on Underwater Wireless Sensor Network and MAC Protocol and Location Algorithm", DOI: 10.1109/ACCESS.2019.2901375, Electronic ISSN: 2169-3536, IEEE.

[5] Judith Santana Abril, Graciela Santana Sosa, and Javier Sosa (2019), "Design of a Wireless Sensor Network for Oceanic Floating Cages in Aquaculture", DOI: 10.1109/MWSCAS.2019.8885256, Electronic ISBN: 978-1-7281-2788-0, IEEE, pp.977-980.

[6] AliyuDala, TughrulArslan, and Imran Saied (2019), "Design of a Triangular Slotted Parasitic Yagi-Uda Antenna for Underwater Linear Sensor Network," IEEE, DOI: 10.1109/comite.2019.8733431, ISBN: 978-1-5386-9337-7.

[7] En Cheng, Longhao Wu, Fei Yuan, Chuanxian Gao, Jinwang Yi (2019), "Node selection algorithm for underwater acoustic sensor network based on particle swarm optimization," IEEE Electronic ISSN: 2169-3536, DOI: 10.1109/ACCESS.2019.2952169.

[8] MohamadMortadaa, AbdallahMakhoula, ChadyAbouJaoudeb, Hassan Harbb, and David Laiymani (2019), "A Distributed Processing Technique for Sensor Data Applied to Underwater Sensor Networks," IEEE, pp.979-984, DOI: 10.1109/IWCMC.2019.8766742, Electronic ISBN: 978-1-5386-7747-6.

[9] Gang Zhao, Yaxu Li, and Lina Zhang (2019)SSEEP: State-Switchable Energy-Conserving Routing Protocol for Heterogeneous Wireless Sensor Networks, DOI: 10.1109/ICEIEC.2019.8784570, Electronic ISBN: 978-1-7281-1190-2, pp.685-689.

[10] ShreemaShetty, Radhika M Pai&Manohara M. M. Pai (2018), "Design and implementation of aquaculture resource planning using underwater sensor wireless network", ISSN: (Print) 2331-1916 (Online) 6 <https://doi.org/10.1080/23311916.2018.1542576>. Cogent Engineering (2018).

[11] NadeemJavaid, HammadMaqsood, Abdul Wadood, IftikharAzimNiaz, Ahmad Almogren, AtifAlamri, and ManzoorIlahi (2017), "A Localization Based Cooperative Routing Protocol for Underwater Wireless Sensor Networks", <https://doi.org/10.1155/2017/7954175>, Hindawi, pp.1-39.

[12] KumuduMunasinghe, Mohammed Aseeri, Sultan Almorqi, Md. FarhadHossain, MusbihaBinteWali, and Abbas Jamalipour (2019), "EM-Based High Speed Wireless Sensor Networks for Underwater Surveillance and Target Tracking", <https://doi.org/10.1155/2017/6731204>, Hindawi Journal of Sensors. Pp.1-14.

[13] YishanSu ,YongpengZuo, Zhigang Jin , and Xiaomei Fu (2019), "OSPG-MAC: An OFDMA-Based Subcarrier Pregrouping MAC Protocol for Underwater Acoustic Wireless Sensor Networks", <https://doi.org/10.1155/2019/4965231>, Article ID 4965231, Hindawi Journal of Sensors, pp.1-12.

[14] Pei-Hsuan Tsai, Rong-Guei Tsai, and Shiuan-Shiang Wang (2017), "Hybrid Localization Approach for Underwater Sensor Networks", <https://doi.org/10.1155/2017/5768651>, Article ID 5768651, Hindawi Journal of Sensors. Pp.1-13.

[15] Fang Zhu and Junfang Wei (2018), "An Energy Efficient Routing Protocol Based on Layers and Unequal Clusters in Underwater Wireless Sensor Networks", Article ID 5835730, <https://doi.org/10.1155/2018/5835730>, Hindawi Journal of Sensors. pp.1-11.

# The Trend of Segmentation for Arabic Handwritten Touching Characters

Ahmed Mansoor Mohsen Algaradi<sup>1</sup>, Mohd Sanusi Azmi<sup>2</sup>, Intan Ermahani A. Jalil<sup>3</sup>

Abdulwahab Fuad Ayyash Hashim<sup>4</sup>, Afrah Abdullah Muhammad Al-Malki<sup>5</sup>

Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka<sup>1,2,3</sup>

Network Systems Support Engineering, College of Telecom and Electronics, Jeddah, Kingdom of Saudi Arabia<sup>4</sup>

Computer Science, Umm Al Qura University, Adhm, Kingdom of Saudi Arabia<sup>5</sup>

**Abstract**—The paper is a comprehensive study of existing research trends in the sector of Arabic language, with a focus on state-of-the-art methods to illustrate the existing condition of various theory in that sector, with the goal of facilitating the adaptation and extension of prior ones into new systems and applications. In the Arabic alphabet, there are 28 letters. Depending on its place in the word, every Arabic letter has over one shape; a single character may have from one to four shapes. The Touching between character and the Overlapping occurred in the handwritten. Historical documents contained a massive knowledge and culture. There are many old books that need to be converted into readable format. Which would take a long time if humans converted it. However, the main problem is the lack of research in Arabic Handwritten especially for segmentation of touching characters. Thus, current trends of the segmentation techniques are investigated to identify the current state-of-the-art of segmenting touching characters in other domains for constructing enhance techniques for Arabic touching characters. In this paper, it reviewed approaches for the segmentation of the touching characters. This paper presents the trend of approaches for the recognition process and segmentation of Arabic handwritten touching characters. In this paper, it highlighted the strength of each technique, the method used, and the drawback of the techniques. Based on the outcome, this will provide a good foundation for constructing a better technique for segmentation of Arabic touching characters, especially from the degraded documents.

**Keywords**—Component; character segmentation; Arabic handwritten; character touching; recognition

## I. INTRODUCTION

Arabic is now the official language of nearly 26 nations, with a population of 280 million people globally. It is among the six official languages of the United Nations (UN) (Chinese, Arabic, English, French, Russian, and Spanish). Furthermore, several of its vocabulary and forms are used in Persian (Farsi), Jawi, Kurdish, Urdu, and Pashto.

Some individuals here nowadays mostly use pen and paper to write notes (for instance). That strategy has a number of flaws. Handwritten text is difficult to retain and access in an efficient and appropriate manner. Searching through them and sharing them with others is a time-consuming process. A lot of critical knowledge may be lost and not utilized efficiently if that content was not available in electronic form.

The segmentation might confront various difficulties. In addition, character should not be too tinny and neatly segmented to better identify the recognition process [1]. The Arabic word is often a line that draws this intricacy of segmentation [2]. Because it used computers in almost every aspect of life, it also known the modern era as the information technology era. The computer is a necessary component of human life. Although, compared to humans, computers do not have nearly as much intelligence. Humans can recognize any sort of text picture from old and deteriorated texts in libraries, but computers cannot comprehend these text images directly [3]. Offline handwritten touching Arabic characters segmentation is a popular topic in study, however it's fraught with difficulties because to differences in writing, overlapping, and touching letters. The segmentation becomes tough when two characters are related to each other [4]. Mostly, all libraries and national archives throughout the world hold large volumes of historical and deteriorating documentation as a book. To convert these important resources to a machine-readable file, special care must be taken [5]. The Arabic language comprises 28 letters, each of which has a distinct form. Because letters in writings are combined to create words, these connections affect the appearance of the letters, thus the shape of an isolated character differs from the shape of a character in the middle and end of the word [6]. Segmentation is closely connected to recognition since it is a highly significant and key phase that splits a picture into sub-units such as lines, words, and letters [7].

OCR (Optical Character Recognition) is a technique that converts scanned or other kinds of pictures into editable format [8]. But even though picture segmentation is not strongly associated with image recognition, the two are inextricably linked. Segmentation process is a critical foundation for image recognition [4]. Picture segmentation, a critical process, splits the picture into tiny pieces.

Even though handwriting is common and varies from person to person, segmentation, which is used to break the text into lines, words, and characters of handwritten text, is still a difficult task. As a result, many observers are going to investigate answers to solve the problem, and some of them have made notable achievements; however, more research is needed to improve the performance of already developed systems. Although it is impossible to explain all the established approaches in this work, the study conducted by addressing the difficulties of touching Arabic handwritten letters [9].

However, this paper aims to show the results and specifications of each segmentation method to assist researchers in determining the best technique for their work.

The rest of the paper is arranged as follows. Section II explains the fundamentals of the Arabic language's characteristics Section III describes the works that are related. Results and discussion details are in Section IV. Section V discusses the conclusion and next work for further study.

## II. RELATED WORK

According to a review of the published literature on the segmentation of touching characters, there is a lack of research effort for handwritten and typed Arabic characters when compared to the number of techniques proposed for other languages such as Chinese and English.

In [13], for printed Arabic text, propose a segmentation based on Omni typeface and open-vocabulary OCR. The APTID-MF dataset was chosen as the basis for the suggested approach. This method does not need an explicit font type identification stage. The method used in this work requires cautious management, since picture samples produced by conventional image augmentation algorithms might lose important features and can be linked.

According to [14], to segment Arabic handwritten text, a region-based approach is used to extract the diacritic. After grayscale the picture, they binarize it, then use the region-based method, and finally extract the diacritic from the image. The researcher utilized the Al Quran as a dataset and added 10 handwritten Arabic pictures. This study also addresses diacritics, which are crucial to the syntax and semantics of a word. While it is part of the alphabet, the points and hamza "ء" are considered as diacritics.

Meanwhile [15] the researcher identifies fork points on handwritten Chinese character skeletons. The primary goal of this study is to increase the proportion of segmentation and recognition. The method identifies the feature point in the binary picture, then thins and smooths the character image to identify the fork and endpoints. Following that, they make some changes to eliminate the erroneous branches. They make use of the DHCCRL database. The rectification of form distortion and the selection of 6,000 handwritten Chinese character pictures are two of the work's highlights.

In addition, [16] method for developing a junction detecting algorithm the researcher omitted a database in this study. This study is just for the Printed Uppercase Alphabet and only between two characters. In this study, just one segmentation instance was investigated. The case presented in this study is neither trustworthy nor practical. In fact, the touching in the writing is more difficult.

Fig. 1 illustrates an instance in which they tested the two characters created by the researcher and placed a straight line between them; normally, it is not touched in this manner during natural handwriting compared to the case shown in the figure below, which is quite significant.

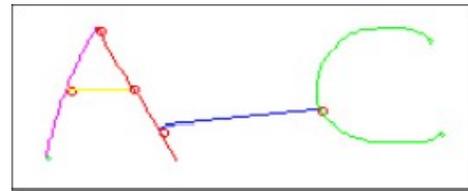


Fig. 1. Example Two Touched Characters [16].

Moreover [17] a technique based on junctions was used to create a handwritten Devanagari character by using a combination of feature to extract the character. Beginning with Handwritten Character Transformation to Bit-mapped Binary Images, the binary image was scaled, and then The Extraction was performed. They get the data from the CVPR Unit, ISI, and Kolkata. One benefit of this research is the collection of 4900 handwritten Devanagari characters. There are five options in this research. On the other side, it might be claimed that there is an advantage to having a lot of options. If, for example, two characters cannot be effectively segmented, the other option can be used.

Furthermore, in [18] Inam Ullah used the junction method while handling Arabic handwritten text. The picture is transformed to binary, and just one point of thickness is kept, making it easier to discern endpoints. However, the intersection set theory is then applied to determine the junction point and broken character. The major goal of this research is to use the algorithm to convert a handwritten, unreadable old Arabic book into a readable one. One advantage of this study is that identifying the endpoints aids in the discovery of the broken character. The researcher chooses the contact point by hand from one of four datasets: IFN/ENIT, CEDAR, and IFN/ENIT, Arabic Dataset, AHDB, and Arabic Handwritten 1.0.

In [19], segmentation of Arabic handwritten text has been performed using contour analysis. In this research, the page is divided into lines initially. Second, the line is divided into sub-words, and last, the sub-words are divided into characters. This method makes use of the database IFN/ENIT. Instead of identifying the baseline or intersecting points, this study replicates the human analogue in Arabic text writing.

Likewise, in Inam Ullah [9], the touching Arabic handwritten characters were segmented using contour tracing. Remove unnecessary noise from a binary picture. Identifying the End, Touching, and Neighboring Points Direction should be written. In the end, they are divided into characters. Many databases were considered, including AHDB, IFN/ENIT, Arabic handwritten 1.0, IBN SINA, IAM, and NIST. Because of proper segmentation, this study could achieve 97.27 percent.

Referring to [20] Corner detection in pictures is a fundamental computer vision problem.

In Lamia Berriche (2020) [24] the technique used is Seam carving-based and Datasets are IESK-ArDB and IFN/ENIT this method leads to Result of 95.67% clear remark for this research is that small characters could be considered secondary components.

Finally, according to [5], the researcher ran one set of 100 words without overlapping and another set of 100 words with overlapping from the benchmark database. And next apply the Method on the handwritten words and report the results for only the second batch. As it stated, it is a simple method that is straightforward to use and quick. Slant correction approaches do not give good results when writing characters with severely slanted and horizontally overlapping characters. Few letters, such as u, v, w, m, and n, are over-segmented or skipped segmented. In Core-zone detection, the researcher advised to count the white pixel until the first major change happens. But how can determine if this one is significant or not? This is a fluid word, and anyone may argue for or against it. Because science only speaks the language of numbers. Their method is straightforward; however, they cannot provide the results of segmentation before and after using the Core-zone detection. As a result, it can be determined if it is essential or not. The researcher simply stated that the first set of words is excellent, with no percentage showing how much is good, so that may compare the overlapping and non-overlapping sets. Also, make it more dependable.

### III. ARABIC LANGUAGE CHARACTERISTICS

#### A. Location/Direction in Writing

In both handwritten papers and machine printed materials, Arabic text is written from right to left, but numerals are written in the same way as numbers in other languages, for example, from left to right [10], [11].

Fig. 2 shows one example.



Fig. 2. Direction for Arabic Writing.

Fig. 3 shows an Arabic numeral example.



Fig. 3. Arabic Number Direction.

#### B. Shape of Arabic Characters

Because Arabic writing letters are interconnected with each other, virtually every character in the Arabic language changes its shape in writing in word according to its placement in the word.

Fig. 4 depicts Arabic letters that change shape depending on their location in an Arabic word, as well as instances of how these characters are linked to form words. The picture illustrates four Arabic letters as an example. However, not much different in the rest of the Arabic language letters regarding the shape forms of the letters compared to the selected four alphabets [12].

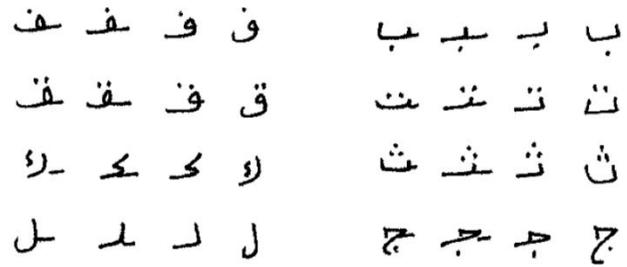


Fig. 4. Example of Shape Alphabet [12].

### IV. CHALLENGES AND LIMITATIONS

There are several obstacles for academics to address in this field, and there is a desire for new ways to develop as computer technology improves and resource constraints diminish [26].

Based on Ouwayed and Belaid's study [23], Kang [22], Aouadi [21], and Saber et al. developed a method for segmenting touching Arabic letters in the same word or other words on the same line or other lines. These existing approaches are template-based segmentation techniques, in which a glossary file is created for all potential touching graphics, that is not only time-consuming due to the variation in Arabic writing and similarities in Arabic characters, but it also fails to address the issue of touching Arabic handwritten characters. Whereas these approaches employed self-defined criteria to govern segmentation accuracy, the segmentation process of touching character pictures suffered as a result.

Over or under segmentation happens because of datasets utilized, languages type (since Arabic has more issues than other languages), type of data (printed or written by hand), and suggested segmentation technique.

By referring to [25] there were some of the challenges such as: Datasets of Arabic handwritten characters, preprocessing noise, Techniques that are cutting-edge, Documents of low resolution and quality, Segmentation, Systems that operate in real time.

The factor that considered as the main factor which is the segmentation. Certain earlier efforts relied on manually dataset segmentation, while others relied on segmented databases. A few of the accessible datasets are not segmented, while others relied on segmented datasets. It's crucial to find a scalable approach to automatically divide documents into lines and subsequently into words (or characters), particularly for big and ancient datasets. Another difficulty in segmentation is dealing with ligatures and the large quantity of Arabic characters.

The multiple sub-words could affect the segmentation process some of the words with single sub-word such as: "محمد" and it could reach to five sub-words for example: "أوروبا" which could increase the difficulties to recognize it as one word during the segmentation process.

Fig. 5 illustrates the challenges of the existing approaches for Arabic Handwritten touched characters segmentation.

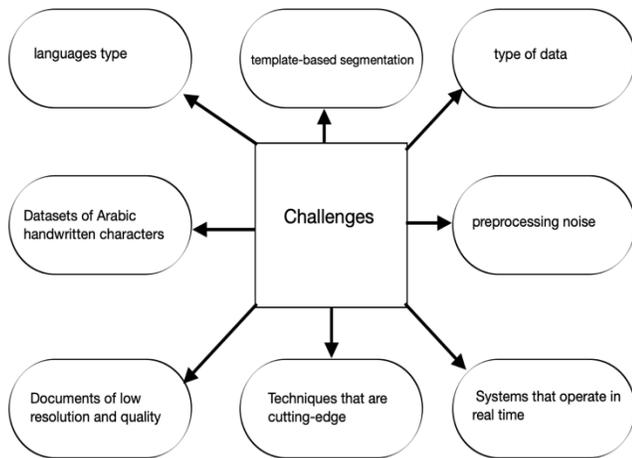


Fig. 5. Challenges of the Existing Approches.

### V. RESULT AND DISCUSSION

After the research method is to find the most successful approach for Arabic handwritten touching character segmentation, but because of the many factors that need to be considered, such as paper quality, number of touching characters that have been tested, database selected, methods used, algorithm applied, and time taken to segment character. Because of all of that, it is difficult to give certain results, especially if some of these factors are not mentioned in the study. However, the author has reviewed ten of the approaches. Table I shows the sample of comparison for each method with its database selected and the result. Author has found a serious need for a specific database which could improve the future research and ease the way for the research to become more reliable, which has a logical result to be compared among the other studies.

TABLE I. METHOD COMPARISON

| Method                          | Methods Comparison                                          |        |
|---------------------------------|-------------------------------------------------------------|--------|
|                                 | Dataset                                                     | Result |
| Seam carving-based              | IESK-ArDB and IFN/ENIT                                      | 95.66% |
| segmentation-based on Omni font | APTID-MF                                                    | 95%    |
| region-based technique          | Al Quran                                                    | 80%    |
| Fork Points on the Skeletons    | DHCCCRL                                                     | 99.41% |
| junction detection algorithm    | -                                                           | 100%   |
| Junction based approach         | CVPR Unit, ISI, and Kolkata                                 | 92.8%  |
| junction approach in Arabic     | IFN/ENIT, CEDAR, AHDB, Arabic Handwritten 1.0.              | 93.3%  |
| contour analysis segmentation   | IFN/ENIT                                                    | 89.4%  |
| Contour Tracing                 | AHDB, IFN/ENIT, Arabic handwritten 1.0, IBN SINA, IAM, NIST | 97.27% |
| Core-Zone                       | CEDAR                                                       | 92.6%  |

The author discovered that the junction algorithm developed by InamUllah yields the highest percentage of segmentation accuracy while being a simple process consisting of three main steps: binary process, thinning process that allows tracing the boundary of the character and if there are more than two binary points, it means there is a junction point to be segmented, and segmentation. However, this study has limitations for future work, such as: during the thinning process, some of the elements may be missing or counted as secondary objects; additionally, the alphabet may be triggered due to its tail. Furthermore, the method could not segment more than one junction point at the same time.

### VI. CONCLUSION

The results of this study revealed current research trends in the field of Arabic. It emphasized the present state of several research elements in that field. This can encourage and make it easier to adapt and extend existing systems to new applications and systems. Arabic has a vast and undiscovered reach; nevertheless, little research has been done in that field previously.

We exhibited some of their prior work that was similar to contemporary state-of-the-art methodologies, with fewer mistakes and a high degree of abstraction. As demonstrated in the difficulties section, this identification is meant to give recommendations for future advancements in the field.

Because of the quality of screening, touching handwritten characters is present in old manuscripts. The Author therefore found that touching characters occurs widely in English, Chinese, Devnagari, Numbers and Arabic handwritten historical materials by exploring the literature for the review. This paper is scanning several approaches to help the researchers in this field to find the advantage and disadvantage of these approaches. For future research, the researcher encourages develop a database for touching characters in Arabic language to give more attention to multiple overlapping.

### ACKNOWLEDGMENT

The author would like to thank his brothers Abdulrahman Algaradi and Abdullah Algaradi for sponsoring him during this research.

### REFERENCES

- [1] Farulla, G. A., Murru, N., & Rossini, R. (2017). A fuzzy approach to segment touching characters. *Expert Systems with Applications*, 88, 1-13.
- [2] I. Ullah, M. S. Azmi, and M. I. Desa, "Junction point detection and identification of Broken character in touching Arabic Handwritten text using overlapping set theory," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 256-260, 2019, doi: 10.14569/ijacsa.2019.0100636.
- [3] S. A. Malik, M. Maqsood, F. Aadil, and M. F. Khan, "An efficient segmentation technique for urdu optical character recognizer (OCR)," in *Lecture Notes in Networks and Systems*, vol. 70, Springer, 2020, pp. 131-141. doi: 10.1007/978-3-030-12385-7\_11.
- [4] Farulla, G. A., Murru, N., & Rossini, R. (2017). A fuzzy approach to segment touching characters. *Expert Systems with Applications*, 88, 1-13.
- [5] Saba, T., Rehman, A., & Zahrani, S. A. (2014). Character segmentation in overlapped script using benchmark database. *Computers, automatic control, signal processing and systems science*, 140-143.

- [6] I. Kacem, P. Laroche, Z. Róka, Institute of Electrical and Electronics Engineers. French Section, O. et M. des S. Université de Lorraine. Laboratoire de Conception, and Institute of Electrical and Electronics Engineers, 2014 International Conference on Control, Decision and Information Technologies (CoDIT): proceedings: Université de Lorraine, France, LCOMS, Metz, November 3-5, 2014.
- [7] A. Lawgali, "A Survey on Arabic Character Recognition," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 8, no. 2, pp. 401–426, Feb. 2015, doi: 10.14257/ijcip.2015.8.2.37.
- [8] N. Vincent and J. M. Ogier, "Shall deep learning be the mandatory future of document analysis problems?," *Pattern Recognition*, vol. 86, pp. 281–289, Feb. 2019, doi: 10.1016/j.patcog.2018.09.010.
- [9] I. Ullah, M. S. Azmi, M. I. Desa, and Y. M. Alomari, "Segmentation of touching Arabic characters in Handwritten documents by overlapping set theory and contour tracing," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 5, pp. 155–160, 2019, doi: 10.14569/ijacsa.2019.0100519.
- [10] Khreisat, L. (2006). Arabic Text Classification Using N-Gram Frequency Statistics A Comparative Study. DMIN, 2006, 78-82.
- [11] A. Amin, "OFF-LINE ARABIC CHARACTER RECOGNITION: THE STATE OF THE ART Arabic characters Off-line recognition Handwriting recognition Segmentation Feature extraction Neural Network classifiers Hidden Markov Models Optical character recognition," 1998.
- [12] Y. Boulid, A. Souhar, and M. Y. Elkettani, "Handwritten Character Recognition Based on the Specificity and the Singularity of the Arabic Language," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, p. 45, 2017, doi: 10.9781/ijimai.2017.446.
- [13] A. Qaroush, A. Awad, M. Modallal, and M. Ziq, "Segmentation-based, omnifont printed Arabic character recognition without font identification," *Journal of King Saud University - Computer and Information Sciences*, 2020, doi: 10.1016/j.jksuci.2020.10.001.
- [14] A. A. Sheikh, M. S. Azmi, M. A. Aziz, M. N. Al-Mhiqani, and S. S. Bafjaish, "Diacritic segmentation technique for Arabic handwritten using region-based," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 778–784, Jan. 2020, doi: 10.11591/ijeecs.v18.i1.pp478-484.
- [15] Liu, K., Huang, Y. S., & Suen, C. Y. (1999). Identification of fork points on the skeletons of handwritten Chinese characters. *IEEE transactions on pattern analysis and machine intelligence*, 21(10), 1095-1100.
- [16] U. K. S. Jayarathna and G. E. M. D. C. Bandara, "A Junction Based Segmentation Algorithm for Offline Handwritten Connected Character Segmentation," 2006.
- [17] IEEE Region 10. Colloquium (3rd: 2008: Indian Institute of Technology Kharagpur), Institute of Electrical and Electronics Engineers. Kharagpur Section., IEEE Sri Lanka Section., and Damodar Valley Corporation., IEEE Region 10 Colloquium and Third International Conference on Industrial and Information Systems: ICIS-2008, December 8-10, 2008: theme: "Real-time communicative intelligence for tomorrow's industry": e-proceedings. IEEE, 2008.
- [18] I. Ullah, M. S. Azmi, and M. I. Desa, "Junction point detection and identification of Broken character in touching Arabic Handwritten text using overlapping set theory," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, pp. 256–260, 2019, doi: 10.14569/ijacsa.2019.0100636.
- [19] Computing, Electrical and Electronics Engineering (ICCEEE), 2013 International Conference on. [publisher not identified], 2013.
- [20] J. Zhang, T. Luo, G. Gao, and L. Lian, "Junction point detection algorithm for SAR image," *International Journal of Antennas and Propagation*, vol. 2013, 2013, doi: 10.1155/2013/357379.
- [21] Aouadi, N., Kacem, A., and Belaiad, A., 2014. Segmentation of touching component in Arabic manuscripts. In *Proceedings of the ICFHR*, 1(4), pp. 452–457.
- [22] Kang, L., Doermann, D. S., Cao, H., Prasad, R., and Natarajan, P., 2012. Local segmentation of touching characters using contour based shape decomposition. In *2012 10th IAPR International Workshop on Document Analysis Systems*, pp. 460–464.
- [23] Ouwayed, N., and Belaïd, A., 2009. Separation of overlapping and touching lines within handwritten arabic documents. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5702 LNCS, pp. 237–244.
- [24] Berriche, L., & Al-Mutairy, A. (2020). Seam carving-based Arabic handwritten sub-word segmentation. *Cogent Engineering*, 7(1), 1769315.
- [25] Balaha, H. M., Ali, H. A., & Badawy, M. (2021). Automatic recognition of handwritten Arabic characters: a comprehensive review. *Neural Computing and Applications*, 33(7), 3011-3034.
- [26] Eikvil, L. (1993). OCR-optical character recognition.

# What Influences Customer's Trust on Online Social Network Sites (SNSs) Sellers?

Ramona Ramli<sup>1</sup>, Asmidar Abu Bakar<sup>2</sup>, Fiza Abdul Rahim<sup>3</sup>

College of Computing and Informatics, Universiti Tenaga Nasional, Malaysia<sup>1,2</sup>

Razak Faculty of Technology and Informatics, Universiti Teknologi Malaysia, Malaysia<sup>3</sup>

Institute of Informatics and Computing Energy, Universiti Tenaga Nasional, Malaysia<sup>1,2,3</sup>

**Abstract**—Customer trust has been recognized as an essential part of the rising trend of social commerce. Lack of trust facilitates the hesitation of customers to shop online or to avoid them completely. Therefore, it is essential to implement and analyze a way of buyer-seller relationship establishment that will improve customers' trust. This paper aims to develop a trust model of Social Network Sites (SNSs) sellers, and to assess the dimensions and criteria that affects customer's trust on Online Social Network Sites (SNSs) sellers by using Analytic Hierarchy Process (AHP) approach. The study was carried out among those who have transactions with Malaysian online SNSs sellers at least every three months. The findings have indicated the top three influencing criteria: recommendation, transaction safety, and rating. This study provides insight into the customers' thoughts about placing trust on online SNSs sellers for selling and purchasing activities.

**Keywords**—Online commerce; trust; social commerce; multi-criteria decision-making

## I. INTRODUCTION

The emergence of social commerce on Social Network Sites (SNSs) has changed the near-constant connectivity that enables online sellers to connect with customers. Its 24/7 connectivity allows online SNSs sellers to produce their content and exchange products or services with other users. On the other side, the customer may connect with other online SNSs sellers for current information on products or services [1]. Various platforms are used to share information about products and services to increase sales volume [2] to build customer's trust in buying and purchasing activities.

Although social commerce has become widespread, certain challenges lead to the lack of trust among customers on online SNSs sellers. In social commerce, trust is referred to a customer's belief to trust a seller's ability, generosity, integrity, and predictableness [3]. The uncertainty of the level of content provided by users and the lack of face-to-face interactions make trust a crucial component of social commerce [4]. Social interactions between customers are believed to increase customer trust in sellers [5]. Many customers avoid making online purchases due to a lack of trust in online platforms [6]. For example, customers' concerns regarding the quality of the information provided by online SNSs sellers make them trust the information provided by other customers more than they trust the online SNSs sellers. This demonstrates the significance of trust in motivating people to purchase online.

Various studies were carried out to identify different criteria that influence customer's trust on purchase intention in social commerce. In [3], various characteristics influencing customers' trust in social commerce include reputation, size, information quality, transaction safety, communication, economic feasibility, and electronic word-of-mouth (E-WoM) referrals. Another study adopted some constructs from the technology acceptance model (TAM) to describe social commerce constructs [4], and show trust positively affects the purchase intention, consistent with many other TAM researches.

Trust and informativeness are suggested in [7] as social network characteristics that affect trust in social commerce. Analysis among Indonesia backpackers discovered that attitude and compatibility are significantly influenced purchase intention [8]. According to the research findings of young people's trust in tourism sites, trust and satisfaction are more important than site design and E-WoM [9]. A recent study in [10] indicates that trust and satisfaction influence repurchase and E-WoM intentions.

Several studies also examine the influencing criteria by concentrating on particular SNSs as the medium of social commerce. A study in [11] discovered that propensity to trust and testimonial were two factors that influence trust in the online purchase through SNSs. Results from a study conducted in Thailand, users of social commerce are more likely to trust social commerce if it provides adequate online environments that include recommendation and referrals, rating and reviews, communication, security issues and E-WoM [12].

An empirical study on Instagram users discovered that perceived benevolence, perceived integrity of online store and key opinion leader endorsement are significant factors explaining customer trust and later influencing purchase intention [13]. In a recent study examining the relationship between social presence and customer relationship quality as measured by customer commitment and loyalty, it was found that in social commerce, social commerce trust mediates the effect of social presence on both commitment and loyalty of customers [14]. A survey among individuals using social commerce services in Korea revealed four factors that influence purchase intention related to social commerce: economy, necessity, reliability and sales promotion [15]. A quantitative method depending upon the sample size is employed in most of the existing studies. On the other hand,

one study finds the criteria through statistical methods and uses BP Neural Network approach to construct a social commerce trust evaluation model [16].

The trust element in purchasing-related decision-making can be seen as a multi-criteria decision-making (MCDM) problem. Problems are formulated and solved in MCDM with the criteria taken into consideration to assess an alternative's performance [17]. Analytic Hierarchy Process (AHP) is an MCDM approach that integrates quantitative and qualitative techniques [18] for handling measurable and intangible criteria. AHP enables subjective evaluation by using experts' judgement to decide the importance of criteria. Based on its importance, AHP determines the criteria that dominate the decision-making process and prioritize them. In decision-making, AHP reflects the analytical thinking of humans, where the assessment is performed in a hierarchical structure. In addition, all criteria to be taken into account will not be viewed as equal but by relative weight in the decision-making process. AHP decreases bias in the decision-making process by consistently reviewing the experts' evaluations.

This paper presents a study that aims to capture additional aspects of social commerce that supplement customers' insight and also encourage them to evaluate online SNSs sellers relatively. The rationale behind this study is to explore in-depth the factors that influence customers' trust. A sample of customers who have transactions with the identified Malaysian online SNSs sellers at least every three months is selected for experimental evaluation. This study also examined the potential of applying the AHP technique achieve two objectives:

- 1) to rank the criteria based on expert evaluation under multiple criteria relevant to this field.
- 2) to obtain the criteria weights by performing pairwise comparisons of importance between the criteria that influence customers' trust on online SNSs sellers and prioritize the influencing criteria prior to purchasing.

## II. DATA AND METHODOLOGY

We proposed a research model for analyzing customers' trust focused on Malaysian society to evaluate online SNSs prior to buying-selling activities. The model is designed to analyze the relevant factors that influence Malaysian customers' trust in order to validate the buyer-seller relationships, especially in the field of SNSs. The primary goal of this empirical research was to determine whether and how customers' trust on online SNSs sellers is obtained before deciding to purchase. The AHP method used in this study is illustrated in Fig. 1 to investigate the criteria that affect the customers' trust in online SNSs sellers.

In this first step, the decision problem should be defined since it drives the whole process on why AHP has to be used. As the traditional rating method is unable to filter out the responses' inconsistency, the AHP methodology employs the consistency test that screen out inconsistent responses. This study identified criteria based on the previous literature, as shown in Table I [19]. These criteria would be expected to collect the necessary information regarding their buying decision for evaluations and comparisons.

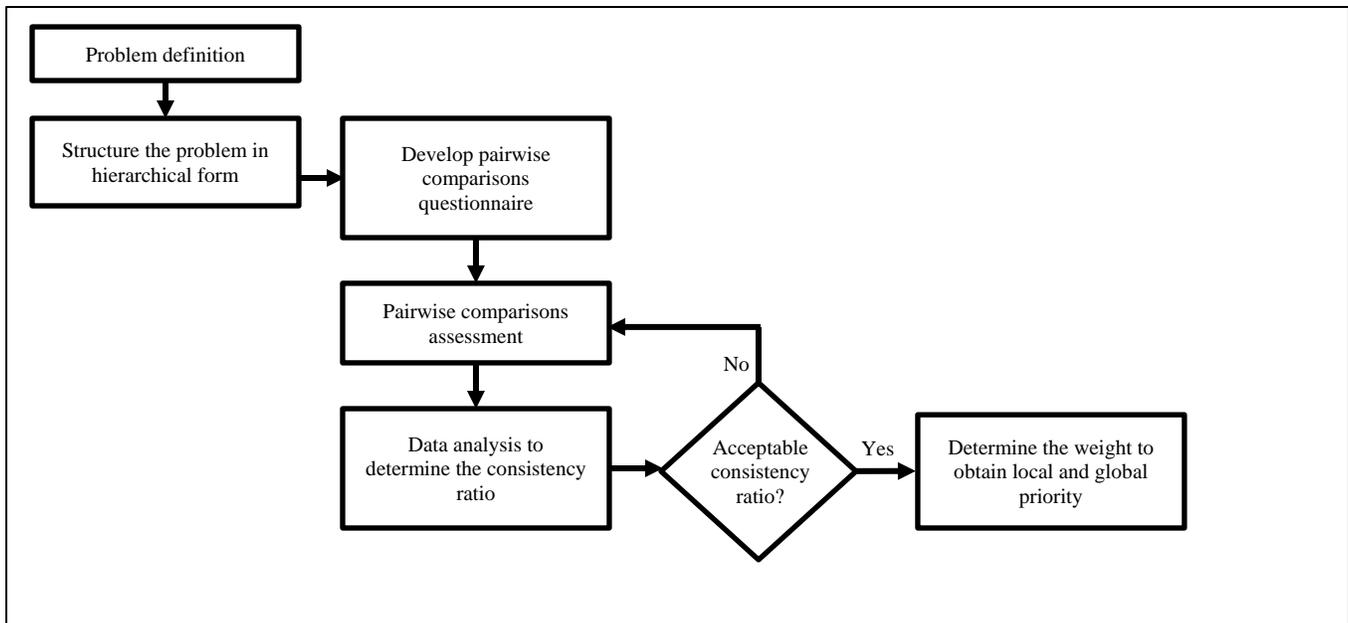


Fig. 1. Overview of AHP Methodology.

TABLE I. PROBLEM IDENTIFICATION

| Dimension                  | Criteria              | References          |
|----------------------------|-----------------------|---------------------|
| E-WoM                      | - Positive Valence    | [3], [4], [20]–[25] |
|                            | - Negative Valence    |                     |
|                            | - E-WoM Content       | [24], [26]–[29]     |
| Information Quality        | - Accuracy            | [3], [30]–[33]      |
|                            | - Relevance           | [30], [32], [33]    |
|                            | - Completeness        | [3], [30]–[34]      |
|                            | - Currency            | [3], [30]–[33]      |
|                            | - Understandability   | [3], [30]           |
|                            | - Format              | [30]–[33]           |
| Social Commerce Constructs | - Recommendation      | [4], [11], [34]     |
|                            | - Rating              | [4], [34]           |
| People                     | - Transaction Safety  | [3], [22], [34]     |
|                            | - Reputation          | [3], [34]           |
|                            | - Propensity to Trust | [22], [34]          |

### III. RESULTS

A schematic representation in a hierarchical structure was formed to structure the problem, as illustrated in Fig. 2. The hierarchy consists of two levels and starts from Level 1 represents the goal, i.e. prioritizing the criteria to evaluate trust on online SNSs sellers. It is then broken up into four dimensions relevant to the goal are represented in Level 2: E-WoM, information quality, social commerce constructs, and people. The sub-criteria associated with each of the dimensions form Level 3.

Once the hierarchical structure has been developed, the relative contribution of each criterion must be obtained through a paired comparison from each expert. In the AHP procedure, the selection of expertise is crucial in establishing the significance of factors in pair comparisons. It is thus vital to identify before a decision has been reached which criteria an expert has to satisfy [35]. Wrong expert selection may lead to a discrepancy in judgement. Furthermore, the qualities required by an expert are varied according to the field of study.

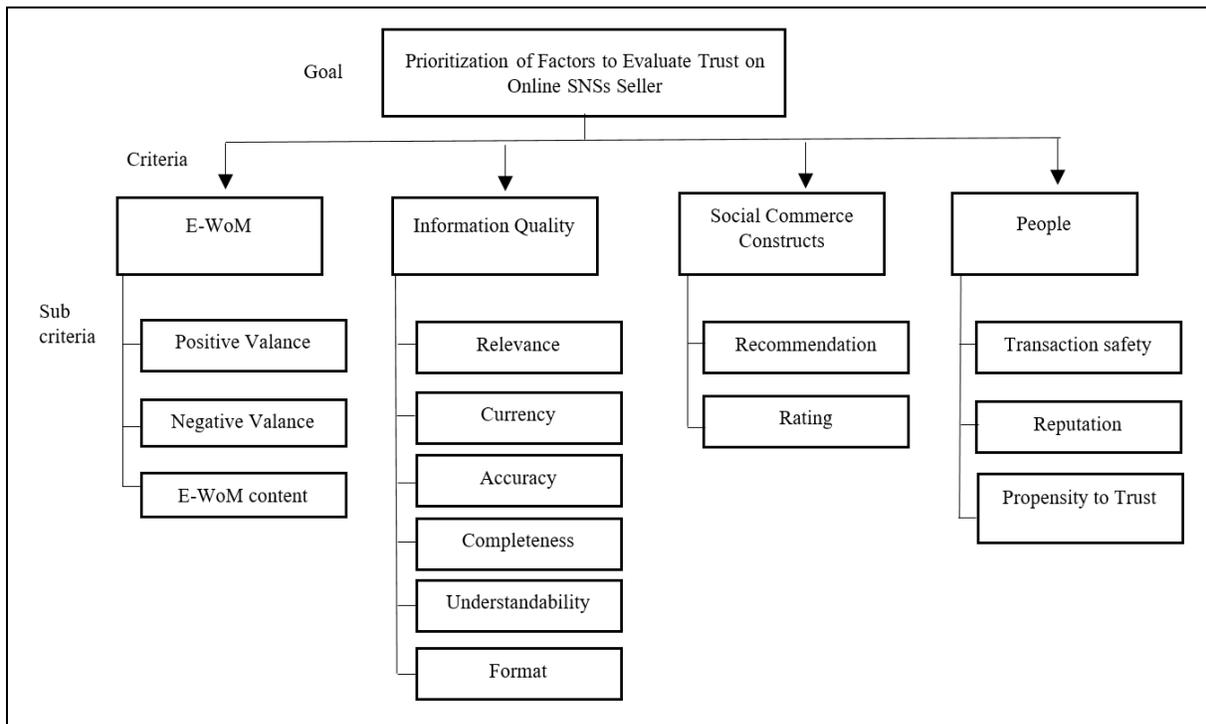


Fig. 2. Hierarchical Structure.

The experts' input plays a significant role in getting the comparison pairwise. In the AHP technique, a single expert is usually assumed to be appropriate to offer a decision. However, a single expert can provide an uncertain judgment. To reduce the uncertain judgment, group decision-making can be employed. The judgment of each member is combined to reach a consensus in group decision-making.

According to Forman and Peniwati [36], two common approaches to aggregate the individual judgments are Aggregation of Individual Judgments (AIJ) and Aggregation of Individual Priorities (AIP). The AIJ combines all individual judgements to act as a 'new' decision by employing geometric mean. The AIP judgment, meanwhile, is based on a number of consensus groups, which may use either arithmetic or geometric mean. All other group decisions are added to the calculated group decisions.

In this study, experts are categorized as people who shop at least once every three months. The definition of experts is based on previously completed similar studies [3], [37]–[42]. 15 respondents took part in making a paired comparison based on the definition of experts, and this can be an excellent contribution to producing a less biased decision.

In order to identify the ranking among factors according to their importance, the information acquired from the questionnaire has been evaluated. The judgement was then consolidated to determine the weight of each criterion by each expert. Criterion ranking was based on the weight determined using the normalized principal Eigen criteria.

The results of this study were based on the evaluations made by the 15 respondents, which were then aggregated using the geometric mean as the AIJ approach was employed. Table II shows a sample of individual evaluations in the form of a pairwise comparison matrix.

The aggregate comparison matrix for criteria for the hierarchical structure level 2 is shown in Table III. An aggregate comparison matrix for sub-criteria in Level 3 to show the relative priorities of the sub-criteria with respect to the Criteria in Level 2 is shown in Table IV (sub-criteria E-WoM), Table V (sub-criteria People), Table VI (sub-criteria Information Quality), and Table VII (sub-criteria Social Commerce Constructs).

TABLE II. SAMPLE INDIVIDUAL JUDGMENT IN THE PAIR-WISE COMPARISON MATRIX

|                   | Accuracy | Relevance | Completeness | Currency | Understandability | Format |
|-------------------|----------|-----------|--------------|----------|-------------------|--------|
| Accuracy          | 1        | 1         | 2            | 1/2      | 1                 | 3      |
| Relevance         | 1        | 1         | 1            | 1        | 2                 | 5      |
| Completeness      | 1/2      | 1         | 1            | 1/3      | 2                 | 3      |
| Currency          | 2        | 1         | 3            | 1        | 4                 | 6      |
| Understandability | 1        | 1/2       | 1/2          | 1/4      | 1                 | 1      |
| Format            | 1/3      | 1/5       | 1/3          | 1/6      | 1                 | 1      |

TABLE III. AGGREGATED COMPARISON MATRIX FOR CRITERIA

|                     | E-WoM | Social Commerce | Information Quality | People |
|---------------------|-------|-----------------|---------------------|--------|
| E-WoM               | 1     | 0.89            | 1.07                | 0.94   |
| Social Commerce     | 1.12  | 1               | 1.25                | 0.94   |
| Information Quality | 0.93  | 0.8             | 1                   | 0.74   |
| People              | 1.06  | 1.06            | 1.35                | 1      |

TABLE IV. AGGREGATED COMPARISON MATRIX FOR SUB-CRITERIA E-WoM

|                  | Positive Valence | Negative Valence | E-WoM Content |
|------------------|------------------|------------------|---------------|
| Positive Valence | 1                | 0.90             | 1.10          |
| Negative Valence | 1.11             | 1                | 1.31          |
| E-WoM Content    | 0.91             | 0.76             | 1             |

TABLE V. AGGREGATED COMPARISON MATRIX FOR SUB-CRITERIA PEOPLE

|                     | Transaction Safety | Reputation | Propensity to Trust |
|---------------------|--------------------|------------|---------------------|
| Transaction Safety  | 1                  | 1.33       | 1.33                |
| Reputation          | 0.75               | 1          | 0.99                |
| Propensity to Trust | 0.75               | 1.01       | 1                   |

TABLE VI. AGGREGATED COMPARISON MATRIX FOR SUB-CRITERIA INFORMATION QUALITY

|                   | Accuracy | Relevance | Completeness | Currency | Understandability | Format |
|-------------------|----------|-----------|--------------|----------|-------------------|--------|
| Accuracy          | 1        | 1.68      | 1.48         | 1.02     | 0.89              | 1.13   |
| Relevance         | 0.60     | 1         | 0.72         | 0.95     | 0.89              | 1.03   |
| Completeness      | 0.68     | 1.39      | 1            | 0.97     | 0.95              | 1.20   |
| Currency          | 0.98     | 1.05      | 1.03         | 1        | 1.03              | 1.37   |
| Understandability | 1.13     | 1.13      | 1.05         | 0.97     | 1                 | 1.28   |
| Format            | 0.88     | 0.97      | 0.83         | 0.73     | 0.78              | 1      |

TABLE VII. AGGREGATED COMPARISON MATRIX FOR SUB-CRITERIA SOCIAL COMMERCE CONSTRUCTS

|                | Recommendation | Rating |
|----------------|----------------|--------|
| Recommendation | 1              | 1.79   |
| Rating         | 0.56           | 1      |

The pairwise comparisons were established for the criteria based on the judgements provided by each expert. The assessment will identify the importance of criteria and sub-

criteria based on the hierarchy structure from step 2. The comparisons are made using a scale of absolute judgements that represents how much one element dominates another with respect to a given attribute.

In determining weights for criteria and sub-criteria, the following procedures are applied to each of the aggregated comparison matrices:

- 1) Calculate the sum values in each matrix column.
- 2) Divide each value in each column by the column sum to normalize the matrix.
- 3) Determine the weight by calculating the average value of each row of the normalized matrix.

The data is analyzed at this point in order to calculate the consistency ratio. If the consistency ratio is not acceptable, the pairwise comparison assessment will be reviewed by experts.

When it comes to making decisions, humans are notoriously inconsistent. The Consistency Ratio (CR) measures the consistency of judgments in order to validate the results. The acceptable CR values (Table VIII) depend on the size of matrices as proposed in [43]. The judgments are consistent and valid if the value is within the range. In this scenario, the experts will be requested to review their judgement. Because of the equation's constraint, the CR value is not relevant to 2X2 matrices [42]. In addition, the two-element matrix has a perfect consistency.

The following four steps are used to compute the CR:

- 1) Multiply each column total by its respective weight for each criteria.
- 2) Get the  $\lambda_{max}$  value by adding values calculated in Step 1.
- 3) Get the Consistency Index (CI) value using Equation (1).

$$CI = (\lambda_{max} - n) / (n - 1) \quad (1)$$

where n is the number of the criteria being compared in the matrix.

- 4) Get the CR value using Equation (2).

$$CR = CI / RI \quad (2)$$

where RI is the Random Index. RI value is determined from a lookup table (Table IX) depends on the n value.

TABLE VIII. ACCEPTABLE CR VALUE

| Size of matrices | Acceptable CR values |
|------------------|----------------------|
| 2x2              | Not applicable       |
| 3x3              | ≤5%                  |
| 4x4              | ≤8%                  |
| Larger           | ≤10%                 |

TABLE IX. RANDOM INDEX

| n | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| R | 0.0 | 0.0 | 0.5 | 0.9 | 1.1 | 1.2 | 1.3 | 1.4 | 1.4 | 1.4 |
| I | 0   | 0   | 8   | 0   | 2   | 4   | 2   | 1   | 5   | 9   |

## IV. DISCUSSION

This section discusses the data analysis for the criteria and each sub-criteria, as well as the determination of weight to obtain local and global priority.

### A. Criteria

People are the most significant criteria, as demonstrated in Table X, with a weight of 28%. Individuals, small or large groups of people, or communities who play a key role in social commerce are referred to as individual consumers and sellers [44].

With a weight of 27%, Social Commerce Constructs is ranked second. Customers can create their own content and share their experiences with online SNSs vendors, products, or services using features offered on social platforms. At the same time, they are permitted to exchange information with others and to offer online social support to other customers. According to [45], social commerce constructs have an impact on customers' trust as well as purchase intention.

In the ranking, E-WoM comes in third. Customers rely on information offered by others to assist them in making purchasing decisions in a virtual world of social commerce. Customers must share any information written by them in order to build confidence with online SNSs sellers.

The Information Quality is ranked last. The customer's absence of direct product experience necessitates adequate, reliable, and high-quality information offered by online SNSs sellers. In addition, the transaction takes place in a non-face-to-face setting.

### B. Sub-Criteria for E-WoM

The ranking of sub-criteria under E-WoM criteria is shown in Table XI. With a weight of 38%, negative valence takes first rank. This is reinforced by a study conducted by [46] which found that negative valence e-WoM had a greater impact on client decision-making than positive valence e-WoM. Positive valence, on the other hand, comes in second rank with a weight of 33%. With a weight of 29 percent, e-WoM material is ranked last. Customers who are heavily involved in online purchases are influenced by the quality of e-WoM, whereas those who are less involved are influenced by the volume of e-WoM [47].

### C. Sub-Criteria for People

The ranking of sub-criteria evaluated under the People criteria is shown in Table XII. With a weight of 40%, Transaction Safety is at the top of the list. Account transfers or bank deposits are the most prevalent payment methods used by customers and online SNSs sellers in social commerce. It is critical for a customer to believe that the seller is protecting their personal and transaction information through this process.

With weights of 30%, Reputation is ranked second. Customers assume that the seller of online SNSs is skilled and trustworthy based on their reputation. Customers will trust online SNSs sellers that have a strong reputation in their buying and selling activities.

TABLE X. NORMALIZED MATRIX AND RANKING WEIGHTS OF CRITERIA

| Criteria                   | E-WoM | Social Commerce Constructs | Information Quality | People | Weights | $\lambda_{max}$ , CI,RI                        | CR     | Ranking |
|----------------------------|-------|----------------------------|---------------------|--------|---------|------------------------------------------------|--------|---------|
| E-WoM                      | 0.243 | 0.237                      | 0.229               | 0.260  | 0.242   | $\lambda_{max}=4.0028$<br>CI= 0.0009<br>RI=0.9 | 0.0010 | 3       |
| Social Commerce Constructs | 0.273 | 0.266                      | 0.268               | 0.260  | 0.267   |                                                |        | 2       |
| Information Quality        | 0.227 | 0.213                      | 0.214               | 0.204  | 0.215   |                                                |        | 4       |
| People                     | 0.258 | 0.283                      | 0.289               | 0.276  | 0.277   |                                                |        | 1       |

TABLE XI. NORMALIZED MATRIX AND RANKING WEIGHTS OF SUB-CRITERIA E-WoM

| Sub-criteria     | Positive Valence | Negative Valence | E-WoM Content | Weights | $\lambda_{max}$ , CI,RI                         | CR     | Ranking |
|------------------|------------------|------------------|---------------|---------|-------------------------------------------------|--------|---------|
| Positive Valence | 0.331            | 0.338            | 0.323         | 0.331   | $\lambda_{max}=3.000$<br>CI = 0.0002<br>RI=0.58 | 0.0004 | 2       |
| Negative Valence | 0.368            | 0.375            | 0.384         | 0.376   |                                                 |        | 1       |
| E-WoM Content    | 0.301            | 0.287            | 0.293         | 0.294   |                                                 |        | 3       |

TABLE XII. NORMALIZED MATRIX AND RANKING WEIGHTS OF SUB-CRITERIA PEOPLE

| Sub-criteria        | Transaction Safety | Reputation | Propensity to Trust | Weights | $\lambda_{max}$ , CI,RI                           | CR       | Ranking |
|---------------------|--------------------|------------|---------------------|---------|---------------------------------------------------|----------|---------|
| Transaction Safety  | 0.399              | 0.398      | 0.401               | 0.399   | $\lambda_{max}=3.000$<br>CI = 5.72E-06<br>RI=0.58 | 9.86E-06 | 1       |
| Reputation          | 0.300              | 0.299      | 0.298               | 0.299   |                                                   |          | 3       |
| Propensity to Trust | 0.300              | 0.302      | 0.301               | 0.301   |                                                   |          | 2       |

TABLE XIII. NORMALIZED MATRIX AND RANKING WEIGHTS OF SUB-CRITERIA INFORMATION QUALITY

| Sub-criteria      | Accuracy | Relevance | Completeness | Currency | Understandability | Format | Weights | $\lambda_{max}$ , CI,RI                          | CR     | Ranking |
|-------------------|----------|-----------|--------------|----------|-------------------|--------|---------|--------------------------------------------------|--------|---------|
| Accuracy          | 0.190    | 0.233     | 0.242        | 0.181    | 0.161             | 0.161  | 0.195   | $\lambda_{max}=6.0444$<br>CI = 0.0088<br>RI=1.24 | 0.0071 | 1       |
| Relevance         | 0.113    | 0.139     | 0.118        | 0.168    | 0.161             | 0.147  | 0.141   |                                                  |        | 5       |
| Completeness      | 0.128    | 0.192     | 0.163        | 0.172    | 0.171             | 0.171  | 0.167   |                                                  |        | 4       |
| Currency          | 0.186    | 0.146     | 0.169        | 0.177    | 0.186             | 0.195  | 0.177   |                                                  |        | 2       |
| Understandability | 0.214    | 0.156     | 0.172        | 0.172    | 0.180             | 0.183  | 0.179   |                                                  |        | 3       |
| Format            | 0.168    | 0.135     | 0.136        | 0.129    | 0.141             | 0.143  | 0.142   |                                                  |        | 6       |

TABLE XIV. NORMALIZED MATRIX AND RANKING WEIGHTS OF SUB-CRITERIA SOCIAL COMMERCE CONSTRUCTS

|                | Recommendation | Rating | Weights | $\lambda_{max}$ , CI,RI     | CR | Ranking |
|----------------|----------------|--------|---------|-----------------------------|----|---------|
| Recommendation | 0.642          | 0.642  | 0.642   | $\lambda=2$<br>CI=0<br>RI=0 | NA | 1       |
| Rating         | 0.358          | 0.358  | 0.358   | 0.168                       |    | 2       |

Propensity is ranked last, with a weight value of the same as Reputation. Since customers and sellers may or may not know each other, trust plays a vital part in social commerce. As a result, each person's level of trust may vary depending on the information available.

#### D. Sub-Criteria for Information Quality

In social commerce, information quality has been identified as a significant criterion in influencing customers' online purchase decisions [48]. The ranking of sub-criteria evaluated under the Information Quality criteria is shown in Table XIII. With a weight of 19 percent, Accuracy is ranked top. Customers may be forced to rely on the information provided

by the seller if they are unable to test the product or services prior to making a purchase. As a result, sellers must be able to deliver information that is accurate, unambiguous, meaningful, believable, and consistent.

With the same weight of 18 percent, Currency and Understandability are ranked second and third. Sellers must present up-to-date information that customers can grasp in order to gain their trust. With a weight of 17 percent, Completeness is ranked fourth. Relevance and Format, meanwhile, are ranked fifth and sixth, respectively, with a 14 percent weighting. The seller's information should represent all conceivable states that are relevant and required by the user when making a purchase choice.

E. Sub-Criteria for Social Commerce Constructs

The ranking of sub-criteria studied under Social Commerce Constructs criteria is shown in Table XIV. Recommendation is ranked first, with a 64 percent weighting. With a weight of 36%, a Rating that represents a measurement scale on a product, service, or seller is ranked second. Customers must rely on other customers' experiences represented through ratings and recommendations because they cannot personally experience the product or services.

F. Determination of Local and Global Weights

The weights for each primary criteria are multiplied by the weights of each relevant sub-criteria to arrive at the global weights. Local weights are the weights assigned to each main and sub-criteria. Table XV reveals that the sub-criteria Social Commerce Constructs' (0.1728) is the most important, followed by Transaction Safety (0.1120) under sub-criteria People. The least important sub-criteria, Relevance and Format of Information Quality, are both weighted at 0.0294.

This study identified new criteria and sub-criteria for evaluating the trustworthiness of online SNSs sellers from a theoretical standpoint. To depict the priority of characteristics that influence customers' trust in those sellers, a hierarchy structural model is built as illustrated in Fig. 3.

Online SNSs sellers can use the identified influencing criteria to improve their business activity in the real world. One

of the most important findings is that online SNSs sellers should pay more attention to customer recommendations, which could lead to increased client trust. Existing customers' positive recommendations may entice new consumers to engage in buying-selling activities. Furthermore, if a known friend recommends an online SNSs sellers, potential clients will have more confidence in the business.

Second, throughout the selling-buying activities, online SNSs sellers must verify that the transaction and its linked information are secure. For transaction payment, the majority of online SNSs sellers employ bank transfers. As a result, it is critical to ensure that the personal information of associated customers is kept secure and not shared with third parties. Customers should also be constantly updated about transaction status and payment confirmation because the transaction does not take place face-to-face.

Finally, customers can use ratings to help them decide whether or not to purchase something. Existing customers rate online SNSs sellers based on their transaction experience. Because a positive rating will help potential clients trust online SNSs sellers, online SNSs sellers must maintain a positive rating in order to expand their business. In addition, customers' trust in online SNS sellers should be based on the People dimension and Social Commerce Constructs as discussed in the analysis results.

TABLE XV. FINAL WEIGHTS AND OVERALL RANKING - GRAPH

| Criteria                   | Local Weights | Sub-criteria        | Local Weights | Global Weights | Overall Ranks |
|----------------------------|---------------|---------------------|---------------|----------------|---------------|
| E-WoM                      | 0.240         | Positive Valence    | 33%           | 7.92%          | 7             |
|                            |               | Negative Valence    | 38%           | 9.12%          | 4             |
|                            |               | E-WoM Content       | 29%           | 6.96%          | 8             |
| Social Commerce Constructs | 27%           | Recommendation      | 64%           | 17.28%         | 1             |
|                            |               | Rating              | 36%           | 9.72%          | 3             |
| Information Quality        | 21%           | Accuracy            | 19%           | 3.99%          | 9             |
|                            |               | Relevance           | 14%           | 2.94%          | 13            |
|                            |               | Completeness        | 17%           | 3.57%          | 12            |
|                            |               | Currency            | 18%           | 3.78%          | 10            |
|                            |               | Understandability   | 18%           | 3.78%          | 11            |
|                            |               | Format              | 14%           | 2.94%          | 14            |
| People                     | 28%           | Transaction Safety  | 40%           | 11.20%         | 2             |
|                            |               | Reputation          | 30%           | 8.40%          | 6             |
|                            |               | Propensity to Trust | 30%           | 8.40%          | 5             |

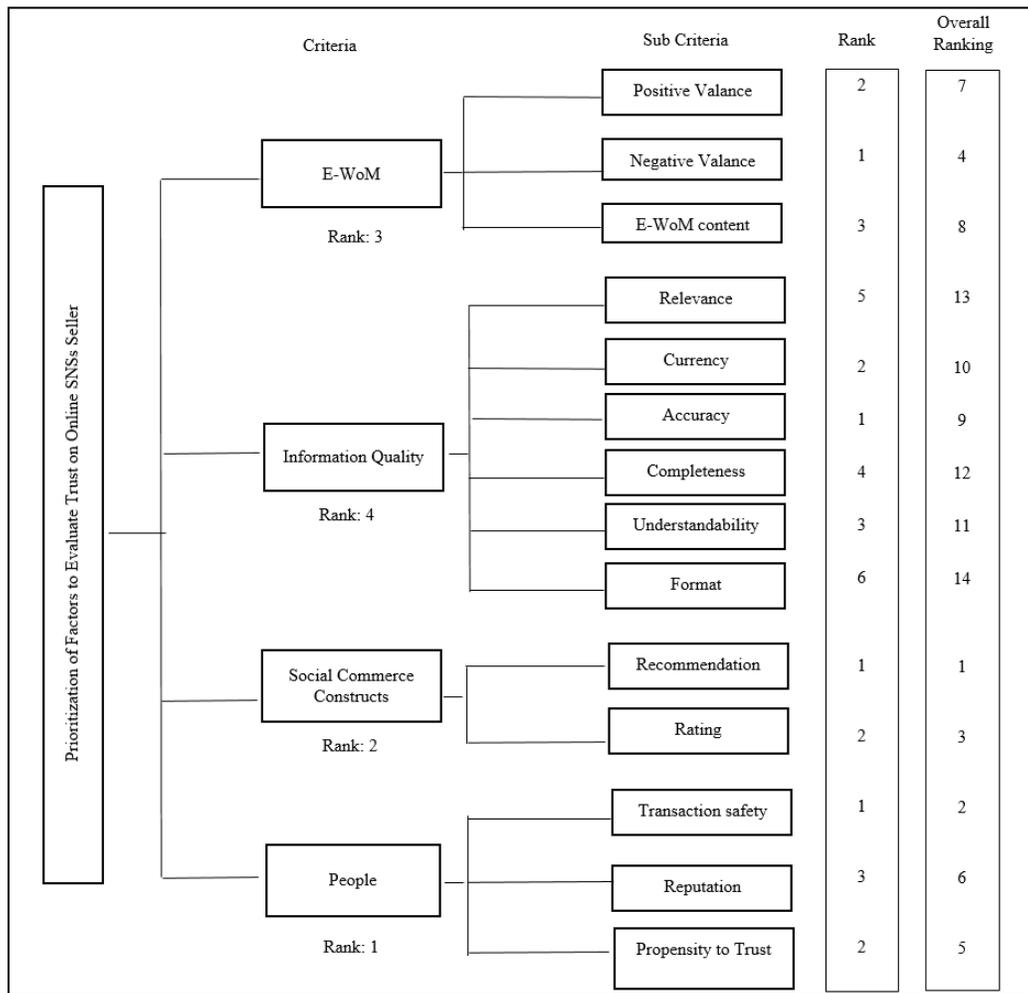


Fig. 3. Hierarchy Structure Model

## V. CONCLUSION

Customers lack direct product experience in social commerce since activities take place in a non-face-to-face environment. A lot of research has been undertaken to investigate the factors that influence customers' trust in online commerce. However, there are few studies that look at the criteria and their significance in the context of online sellers who utilize social media sites as their marketing platforms. This study addresses the gap by examining the factors for boosting customers' trust in online SNSs sellers from many angles.

This study has successfully achieved the objective by using the AHP technique to discover and rank the criteria that influence a customer's trust in online SNSs sellers prior to making a purchase intention. The following criteria are prioritized based on the ranking: (1) Recommendation, (2) Transaction Safety, (3) Rating, (4) Negative Valence, and (5) Propensity to Trust, according to this study. To highlight the priority as well as the interaction between criteria and sub-criteria, a hierarchy structural model is created.

The findings of this study show which criteria the customers should consider according to its importance when

evaluating the reliability of online SNSs sellers prior to making a purchase. The identified criteria, on the other hand, will serve as a guideline for online SNSs sellers to establish customer trust in their buying and selling activities.

The identified criteria and sub-criteria can be investigated further in future research to quantitatively assess the trustworthiness of online SNSs sellers. There may be interrelationships between some of the specified criteria and sub-criteria. In that instance, Analytical Network Process (ANP) might be used to do more research because it ignores interaction between criteria and sub-criteria. The results of this study can be integrated or compared with those from other multi-criteria approaches like TOPSIS, ELECTRE, and SWOT.

## REFERENCES

- [1] T. Ric and D. Benazić, "From social interactivity to buying: an instagram user behaviour based on the S-O-R paradigm," *Econ. Res. Istraživanja*, vol. 0, no. 0, pp. 1–19, 2022, doi: 10.1080/1331677x.2021.2025124.
- [2] L. Guan, H. Chen, H. Ma, and L. Zhang, "Optimal group-buying price strategy considering the information-sharing of the seller and buyers in social e-commerce," *Int. Trans. Oper. Res.*, vol. 29, no. 3, pp. 1769–1790, May 2022, doi: <https://doi.org/10.1111/itor.13075>.

- [3] S. Kim and H. Park, "Effects of various characteristics of social commerce (s-commerce) on consumers' trust and trust performance," *Int. J. Inf. Manage.*, vol. 33, no. 2, pp. 318–332, 2013, doi: 10.1016/j.ijinfomgt.2012.11.006.
- [4] N. Hajli, "Social commerce constructs and consumer's intention to buy," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 183–191, Apr. 2015, doi: 10.1016/j.ijinfomgt.2014.12.005.
- [5] M. N. Hajli, "A study of the impact of social media on consumers," *Int. J. Mark. Res.*, vol. 56, no. 3, pp. 387–404, 2014, doi: 10.2501/UMR-2014-025.
- [6] G. A. Tran and D. Strutton, "Comparing email and SNS users: Investigating e-servicescape, customer reviews, trust, loyalty and E-WOM," *J. Retail. Consum. Serv.*, vol. 53, no. March 2019, p. 101782, 2020, doi: 10.1016/j.jretconser.2019.03.009.
- [7] M.-C. Han, "How Social Network Characteristics Affect Users' Trust and Purchase Intention," *Int. J. Bus. Manag.*, vol. 9, no. 8, Jul. 2014, doi: 10.5539/ijbm.v9n8p122.
- [8] A. P. Aristio, S. Supardi, R. A. Hendrawan, and A. A. Hidayat, "Analysis on purchase intention of Indonesian backpacker in accommodation booking through online travel agent," *Procedia Comput. Sci.*, vol. 161, pp. 885–893, 2019, doi: 10.1016/j.procs.2019.11.196.
- [9] D. Buhalis, E. Parra López, and J. A. Martínez-Gonzalez, "Influence of young consumers' external and internal variables on their e-loyalty to tourism sites," *J. Destin. Mark. Manag.*, vol. 15, p. 100409, Mar. 2020, doi: 10.1016/j.jdmm.2020.100409.
- [10] N. Meilatinova, "Social commerce: Factors affecting customer repurchase and word-of-mouth intentions," *Int. J. Inf. Manage.*, vol. 57, p. 102300, Apr. 2021, doi: 10.1016/j.ijinfomgt.2020.102300.
- [11] K. M. Nor, W. N. Fazni Wan Mohamad Nazarie, and A. A.-A. Md Yusoff, "Factors influencing individuals' trust in online purchase through social networking sites," 2013 7th International Conf. e-Commerce Dev. Ctries. With Focus e-Security, ECDC 2013, 2013, doi: 10.1109/ECDC.2013.6556752.
- [12] C. Pothong and C. Sathitwiriawong, "Factors of s-commerce influencing trust & purchase intention," 20th Int. Comput. Sci. Eng. Conf. Smart Ubiquitous Comput. Knowledge, ICSEC 2016, pp. 0–4, 2017, doi: 10.1109/ICSEC.2016.7859879.
- [13] J. W. S. Che and C. M. K. Cheung, "Consumer Purchase Decision in Instagram Stores : The Role of Consumer Trust," pp. 24–33, 2017.
- [14] W. Nadeem, A. H. Khani, C. D. Schultz, N. A. Adam, R. W. Attar, and N. Hajli, "How social presence drives commitment and loyalty with online brand communities? the role of social commerce trust," *J. Retail. Consum. Serv.*, vol. 55, no. March, p. 102136, 2020, doi: 10.1016/j.jretconser.2020.102136.
- [15] J. W. Sohn and J. K. Kim, "Factors that influence purchase intentions in social commerce," *Technol. Soc.*, vol. 63, p. 101365, Nov. 2020, doi: 10.1016/j.techsoc.2020.101365.
- [16] L. Chen and R. Wang, "A Trust Evaluation Model for Social Commerce Based on BP Neural Network," *J. Data Anal. Inf. Process.*, vol. 04, no. 04, pp. 147–158, 2016, doi: 10.4236/jdaip.2016.44013.
- [17] G. Pasi, M. Viviani, and A. Carton, "A Multi-Criteria Decision Making approach based on the Choquet integral for assessing the credibility of User-Generated Content," *Inf. Sci. (Ny)*, vol. 503, pp. 574–588, Nov. 2019, doi: 10.1016/j.ins.2019.07.037.
- [18] N. Bhushan and K. Rai, *Strategic Decision Making: Applying the Analytic Hierarchy Process*. London: Springer-Verlag London, 2004.
- [19] R. Ramli, A. A. Bakar, and R. Ismail, "The Trust Effect Towards Online Seller in Social Commerce," *Proc. 6th Int. Conf. Comput. Informatics*, no. 030, pp. 317–322, 2017.
- [20] T. Hennig-Thurau, K. P. Gwinner, G. Walsh, and D. D. Gremler, "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?," *J. Interact. Mark.*, vol. 18, no. 1, pp. 38–52, 2004, doi: 10.1002/dir.10073.
- [21] A. Leerapong and A. Mardjo, "Trust and Risk in Purchase Intention through Online Social Network: A Focus Group Study of Facebook in Thailand," *Inf. Manag. Bus. Rev.*, vol. 5, no. 3, pp. 144–154, 2013, doi: 10.7763/JOEBM.2013.V1.68.
- [22] A. D. Noor, R. Sulaiman, and A. A. Bakar, "A Review of Factors that Influenced Online Trust in Social Commerce," *Int. Conf. Inf. Technol. Multimed.*, pp. 118–123, 2014.
- [23] C. S. P. Ng, "Intention to purchase on social commerce websites across cultures: A cross-regional study," *Inf. Manag.*, vol. 50, no. 8, pp. 609–620, 2013, doi: 10.1016/j.im.2013.08.002.
- [24] Y. Wang and C. Yu, "Social interaction-based consumer decision-making model in social commerce: The role of word of mouth and observational learning," *Int. J. Inf. Manage.*, vol. 37, no. 3, pp. 179–189, 2017, doi: 10.1016/j.ijinfomgt.2015.11.005.
- [25] Y. Wang, S. Wang, Y. Fang, and P. Y. K. Chau, "Store survival in online marketplace: An empirical investigation," *Decis. Support Syst.*, vol. 56, no. 1, pp. 482–493, 2013, doi: 10.1016/j.dss.2012.11.005.
- [26] T. Hennig-Thurau et al., "The Impact of New Media on Customer Relationships," *J. Serv. Res.*, vol. 13, no. 3, pp. 311–330, 2010, doi: 10.1177/1094670510375460.
- [27] C. Lin, Y.-S. Wu, and J.-C. V. Chen, "Electronic Word-of-Mouth: The Moderating Roles of Product Involvement and Brand Image," *Proc. 2013 Int. Conf. Technol. Innov. Ind. Manag.*, pp. 29–47, 2013.
- [28] K. Vimaladevi, "A Study on the Effects of Online Consumer Reviews on Purchasing Decision," *Prestig. Int. J. Manag. IT-Sanchayan*, vol. 1, no. 1, pp. 91–99, 2012.
- [29] A. P. a Yayli and I. M. Bayram, "eWOM: THE EFFECTS OF ONLINE CONSUMER REVIEWS ON PURCHASING DECISION OF ELECTRONIC GOODS," *Marketingtrendscongresscom*, 2009.
- [30] J. V. Chen, B. Su, and A. E. Widjaja, "Facebook C2C social commerce: A study of online impulse buying," *Decis. Support Syst.*, vol. 83, pp. 57–69, Mar. 2016, doi: 10.1016/j.dss.2015.12.008.
- [31] R. R. Nelson, P. A. Todd, and B. H. Wixom, "Antecedents of Information and System Quality: An Empirical Examination Within the Context of Data Warehousing," *J. Manag. Inf. Syst.*, vol. 21, no. 4, pp. 199–235, 2005, doi: 10.1362/026725705774538390.
- [32] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations," *Commun. ACM*, vol. 39, no. 11, pp. 86–95, Nov. 1996, doi: 10.1145/240455.240479.
- [33] N. Au, E. W. T. Ngai, and T. C. E. Cheng, "Extending the Understanding of End User Information Systems Satisfaction Formation: An Equitable Needs Fulfillment Model Approach," *MIS Q.*, vol. 32, no. 1, pp. 43–66, 2008, doi: 10.2307/25148828.
- [34] A. A. Syuhada and W. Gambett, "Online Marketplace for Indonesian Micro Small and Medium Enterprises based on Social Media," *Procedia Technol.*, vol. 11, no. Iccci, pp. 446–454, 2013, doi: 10.1016/j.procty.2013.12.214.
- [35] R. Gawlik, M. Głuszak, and A. Małkowska, "The Measurement of Housing Preferences in the Analytic Hierarchy Process," *Folia Oeconomica Stetin.*, vol. 17, no. 1, Jan. 2017, doi: 10.1515/foli-2017-0003.
- [36] Ernest Forman and Kirti Peniwati, "Aggregating individual judgments and priorities with the Analytic Hierarchy Process," *Eur. J. Oper. Res.*, vol. 108, pp. 165–169, 1998.
- [37] K. M. Nor, W. N. Fazni Wan Mohamad Nazarie, and A. A.-A. Md Yusoff, "Factors influencing individuals' trust in online purchase through social networking sites," in 7th International Conference on e-Commerce in Developing Countries: with focus on e-Security, Apr. 2013, vol. 11, no. SPL.ISS., pp. 1–18, doi: 10.1109/ECDC.2013.6556752.
- [38] S. Sukrat, B. Papisatorn, and V. Chongsuphajaisiddhi, "Impact of Customer Trust on Purchase Intention in Organic Rice through Facebook : A Pilot Study," 10th Int. Conf. E-bus., 2015.
- [39] Z. Zamrudi, I. Suyadi, and Y. Abdillah, "the Effect of Social Commerce Construct and Brand Image on Consumer Trust and Purchase Intention," *Profit J. Adm. Bisnis*, vol. 10, no. 1, pp. 1–13, 2016, doi: 10.9876/VOL1ISSN1978-743X.
- [40] ASEAN UP, "Insights and trends of e-commerce in Malaysia - ASEAN UP," 2017.
- [41] N. A. Hashim, S. M. Nor, and H. Janor, "Riding the Waves of Social Commerce : An Empirical Study of Malaysian Entrepreneurs," *Geogr. Online TM Malaysian J. Soc. Sp.*, vol. 2, no. 2, pp. 83–94, 2016.

- [42] M. Dashti, A. Sanayei, H. R. Dolatabadi, and M. H. Moshrefjavadi, "An Analysis of Factors Affecting Intention to Purchase Products and Services in Social Commerce," *Mod. Appl. Sci.*, vol. 10, no. 12, p. 98, 2016, doi: 10.5539/mas.v10n12p98.
- [43] T. L. Saaty, "How to Make a Decision: The Analytic Hierarchy Process," *Interfaces (Providence)*, 1994, doi: 10.1287/inte.24.6.19.
- [44] C. Wang and P. Zhang, "The Evolution of Social Commerce: The People, Management, Technology, and Information Dimensions," 2012.
- [45] N. Hajli and J. Sims, "Social commerce: The transfer of power from sellers to buyers," *Technol. Forecast. Soc. Change*, vol. 94, no. March, pp. 350–358, 2015, doi: 10.1016/j.techfore.2015.01.012.
- [46] T. Hennig-Thurau and G. Walsh, "Electronic Word-of-Mouth: Motives for and Consequences of Reading Customer Articulations on the Internet," ... *J. Electron. Commer.*, vol. 8, no. 2, pp. 51–74, 2003, doi: 10.1504/IJECRM.2008.020411.
- [47] S.-H. Lee, "How do online reviews affect purchasing intention?," *African J. Bus. Manag.*, vol. 3, no. 10, pp. 576–581, 2009, doi: 10.5897/AJBM09.204.
- [48] B. Shen, D. Liu, and L. Tai, "Customer Information Sharing in Social Commerce Based on FIRE Model: The Role of Trust Propensity," 2014 *Int. Conf. Manag. e-Commerce e-Government*, pp. 119–123, 2014, doi: 10.1109/ICMeCG.2014.33.

# New Textual Authentication Method to Resistant Shoulder-Surfing Attack

Islam Abdalla Mohamed Abass<sup>1</sup>, Loay F.Hussein<sup>2</sup>, Tarak kallel<sup>3</sup>, Anis Ben Aissa<sup>4</sup>

Department of Computer Science, Jouf University<sup>1,2,4</sup>

Department of Physics, Jouf University<sup>3</sup>

**Abstract**—Using textual passwords suffer from the balance between security and usability. Password policies are usually adopted by system administrators to force users to choose strong passwords. However, users often use a simple password to make it easy to remember, which reduces the password strength and make it vulnerable to information security threats. When users enter their passwords in public places like airports or cafes, they become exposed to shoulder surfing attacks which are considered as a kind of social engineering. With a little effort, an attacker can capture a password by recording the individual's authentication session or by direct observation. To overcome this vulnerability, we propose a new textual-password approach that uses camouflage characters and a virtual keyboard which leads to generating strong and easy to remember passwords. The perspective of usability and security was evaluated by experimental studies conducted with 65 users and then compared with recent studies. The results showed that the proposed technique has the lowest shoulder surfing success rate with just 3.63% with reasonable usability.

**Keywords**—Shoulder surfing; caesar cipher; virtual keyboard; graphical password; social engineering

## I. INTRODUCTION

Current authentication systems have a lot of weaknesses even if the system is secured, an individual's behaviour may cause a security breach. Users usually pick a short or easy password to remember, which makes their accounts vulnerable to attack and easily guessable. On the other hand, longer passwords are harder to memorise and to type correctly [1][2]. To consider a password as strong it must have eight characters or more, contain numbers, special characters mixed with small and capitalize alphabets [3].

By adding the human factor to the equation of security and how easily social engineering can manipulate the user, textual passwords become vulnerable to spyware attacks, keyloggers, dictionary attacks, and shoulder surfing attacks [4]. Most individuals are aware of security threats but they insist to avoid them. A survey reported that 90% of 152 computer users leaked their passwords. In this situation, forcing the user to create a password according to strict policies will not solve this issue [5]. To overcome the limitations of text-based passwords, many techniques such as two-factor authentication and graphical passwords are used [6]. Moreover, using input devices such as the mouse and touch-screen makes graphical authentication techniques possible. Unfortunately, they are unsecured to many attacks such as shoulder-surfing, spyware, Social Engineering and Dictionary attacks.

Shoulder surfing attack is a type of identity theft, it occurs when the attacker looks over someone's shoulder to get passwords, login PINs or other sensitive personal data. This attack can also be done by a small wireless camera that is easy to install. To overcome this problem, a wide range of research efforts have been done on eye-tracking algorithms. Systems login that is based on gazing to select a character from an on-screen keyboard is one of the solutions for shoulder-surfing but it may take a long entry time and lack of input accuracy [7]. Another approach to solve this problem is using a graphical password or integrating both graphical and textual passwords [8][9].

Graphical password has been widely used, especially on smartphones. An Individual can unlock his smartphone after the correct pattern is mapped out on a three-by-three rectangle, as in Fig. 1. As shown in Fig. 2 all the authentication Graphical methods can be grouped into three categories:

1) *Recognition-based system*: in this method users can login by choosing the correct photo from a list at the signup time [10].

2) *Hybrid system*: this method combines more than one schema to eliminate their issues and produces a more stable and useful system [11].

3) *Recall-Based system*: this method asks the user to draw or write something to use it as the Authentication code. There are two types of Recall base systems:

a) The Pure Recall method is relayed on the user to give the right authentication code at the login time [12].

b) Cued Recall method is based on helping the users to login by giving them a hint to remember their passwords [12].

However, most login graphical techniques may suffer from a complicated algorithm or need a special device, which increases the need to develop a new secure login method.

In this paper, we first discuss the shoulder surfing attack as a part of social engineering, then propose a defensive model that can resist shoulder surfing attacks. The defensive model is designed on the concept that the user can enter random camouflage characters by using a virtual keyboard. The character of the virtual keyboard is shifted by using Caesar cipher which is used to encrypt and decrypt the user input. After applying an experimental study to test the defensive model with 65 participants and analyzing the data, the conclusion is presented to summarize the primary outcomes and to determine model usability and efficiency against shoulder surfing attacks.

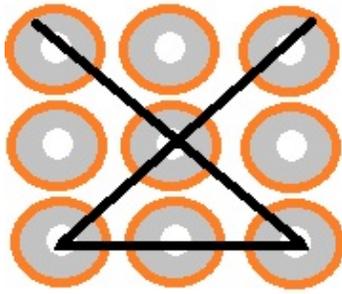


Fig. 1. Touchpoints Pattern to Unlock Smartphone.

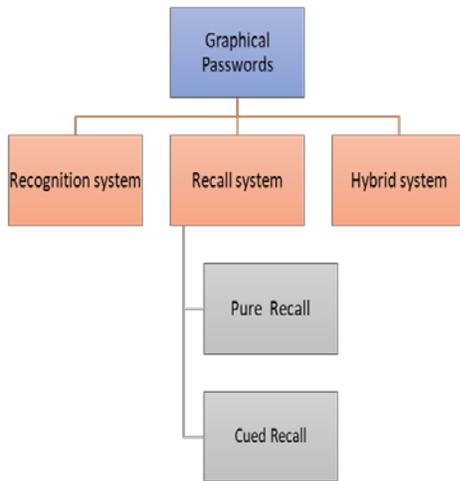


Fig. 2. Categorization of Authentication Methods for the Graphical Password.

## II. RELATED WORK

Graphical password authentication systems that depend on recognition based and recall based schema has been adopted in many research to fight against shoulder surfing attacks. In [13], Jiya Gloria Kaka et al. compared 10 graphical authentication methods based on three common attacks and the usability features. Although, most of the authentication systems have acceptable usability only three methods were effective against shoulder surfing attacks.

In [14], Jianwei Lai et al. introduce a unique authentication scheme that resists shoulder-surfing attacks. The scheme is based on textual passwords, so to login, the user will enter part of the password and skip the password character that is marked with 'x'.

Aakansha S. Gokhale et al. developed a new graphical password authentication technique that resists shoulder surfing, brute force and guessing attacks [15]. The authentication technique is based on three questions related to 25 pictures. In the login phase, the user must click on the correct location in the image for every question. The system has a very large password space equal to  $8.367939e+34$  which can provide a strong secure password.

In [16], Dongmin Choi et al. discussed five model's authentication schemes against shoulder surfing and social engineering attacks. The paper compared defensive types; QWERTY based Secure Keypad, ABC based Secure Keypad,

Touch and Slide Secure Keypad, colour-based Secure Keypad and Random Secure Keypad. As a result, the five techniques were weak against shoulder surfing attacks.

Aravinda Thejas Chandra et al. developed an authentication system based on eye-tracking and a smart camera [17]. The gaze-based PIN identification application was tested on a nine-digits keypad by using real-time eye-tracking to login. Although it can be used as a defensive mechanism against shoulder surfing, the paper did not study the medical effect of the system on the user's eyes for a long period of using special if the password is more than 6 characters.

In [18], Anindya Maiti et al. proposed a defensive model based on a random alignment keyboard with twenty-six alphabets created by an augmented reality wearable device. The model involves three different randomization strategies, each one of them can shift the 26 characters to produce a new virtual keyboard. Although the system is effective as a defensive mechanism against side-channel and shoulder surfing attacks, it requires a secured wireless channel and special hardware.

Eiji Hayashi et al. designed a novel secure mechanism for user authentication that can be used with any screen size [19]. In the proposed model the user chooses a set of images as a graphical password. To login, he must choose his distorted images from the set of images. This authentication technique relies on the fact that human perception is influenced by his information. Despite the simplicity of the method used, it is effective against social engineering and shoulder surfing attacks.

Vishal Kolhe et al. introduced an authentication system based on a 3D password [20]. The 3D password is a multi-passwords system that combines a textual password with a graphical password in a virtual environment. Users can move in a virtual environment to create their passwords. Although the model provides secure authentication and user friendly, it has many disadvantages like time, memory requirement and cost. The system provides immunity against brute force attacks and key loggers but is still vulnerable to shoulder surfing attacks.

Hung-Min Sun et al. introduced a graphical authentication system for smartphones called PassMatrix [21]. The system consists of many components:

- Password verification module.
- vertical and Horizontal axis control module.
- Login indicator generator module.
- Image discretization module.
- Communication module.
- Database.

The images which are used to login are divided into horizontal (1-11) and vertical axis (A-G) grids. To login, the user must circle his hand on the screen to get the generated key or listen to an audio that contains the generated key by using earbuds or a Bluetooth headset. The audio is sent from the

server using a secure channel, then the user must shift the character to match the key. This procedure is repeated until the login is finished. The total accuracy of all login trials is 93:33% which means it can be used to defend against smudge attacks and shoulder surfing attacks.

### III. SOCIAL ENGINEERING AND SHOULDER SURFING ATTACKS

Social engineering is an attack that depends on the human factor. It's classified as a non-technical attack in general. However, it can be combined with technical types of attacks like Trojan and spyware, which makes it more effective [22]. Cyence, a cyber security analyst company reported In 2016, that the United States was the most targeted country with social engineering attacks and had the highest attack cost, followed by Japan and Germany. The total cost of these attacks in the US alone was \$121.22 billion [23].

Shoulder surfing is a kind of social engineering. It can be performed by using technical and non-technical ways. Although shoulder surfing attack can be noticed by the victim, it could be combined with other types of attacks like spear-phishing and dumpster diving which makes it a powerful attack. A survey was done in the US, Germany, and Egypt concluded that 67.4% of Shoulder surfing occur in public places and 74.1% of the observer were strangers [24]. Depending on the Humans nature people act like that for different reasons such as curiosity, boredom or to get private data. Some companies produce special screens that make it difficult to see the smartphone or computer screen from an angle to keep the privacy [25]. This screen can be useless when the attacker uses a small hidden camera that can capture video and send it directly to the attacker. Although there are multi-solutions to solve the shoulder surfing problem, they are suffering from common issues such as:

- 1) The most common software and operating system use alphanumeric usernames and passwords for authentication, which makes it difficult to apply random security algorithms.
- 2) Some graphical password Algorithms used in systems are vulnerable to shoulder surfing and other types of attacks [26].
- 3) Adding layers of security or complex authentication systems decreases the usability and functionality of the system.
- 4) The common careless behaviours like writing down a password or using the data autofill technique make security solutions worthless.

### IV. PROPOSED DEFENSE MODEL

In this study, two different models were designed, a shoulder surfing defensive model and a traditional login system that contains just a password with eight characters and a username. The proposed defensive model is based on camouflage characters and a virtual keyboard with shifted alphabetic as shown in Fig. 3.

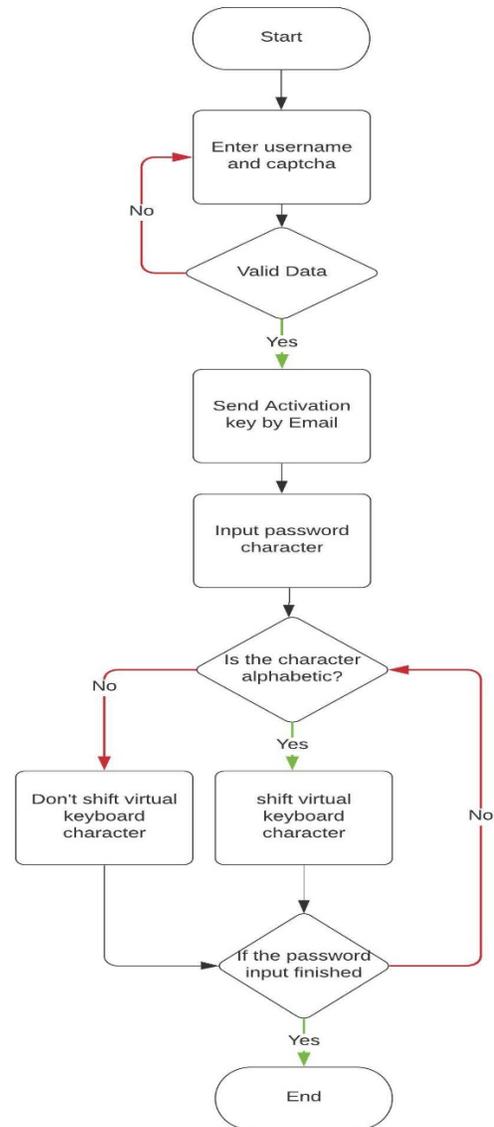


Fig. 3. The General Authentication Mechanism.

The virtual keyboard uses Caesar cipher to shift the 26 alphabetic, while the rest of the keyboard keys which are coloured with blue will be static and will act as normal keys as in Fig. 4. Caesar cipher is a simple substitution cipher that uses the same key for encryption and decryption [27]. Depending on the symmetric key letters are shifted a certain number of places down the alphabet. If the shift exceeds the number of the alphabet, the alphabet will just be rotated to the front. The alphabet in the proposed model is arranged in horizontal lines according to the QWERTY keyboard, so the first alphabetic is q and the last is m. The result of Caesar cipher will be according to the QWERTY keyboard not to the normal alphabetic sequence. This system also applies another layer of protection by adding camouflage characters to the real password. The user authentication process is divided into two phases as follows.



Fig. 4. The Virtual QWERTY Keyboard.

### A. Registration Phase

1) The user must enter eight characters password, username, and email to register an account as in Fig. 5.

2) The traditional system will just use the password and the username to login as in Fig. 6.

### B. Login Phase

1) For authentication, the user must enter his username and captcha character as in Fig. 7. After that, the server will send the activation key to the users' email. The activation key (AK) will be random from 1 to 9 and it mustn't be in the password that has already been entered by the user in the registration phase.

2) In the Caesar encryption algorithm, the encryption key (k) will be chosen randomly from 1 to 10 to shift the character in the virtual keyboard. In the encyusting process, x is the character number in the virtual keyboard and k will be fixed in each authentication session as in Eq (1).

$$En(x) = (x-k) \text{ mod } 26 \quad (1)$$

3) Every time the user enters an alphabetic character, the virtual keyboard will shift only the alphabetic character's position according to the Caesar cipher. Note that the virtual keyboard will shift after pressing alphabetic keys and will not shift, if the user enters a number, special symbols or press any other button. The user password must be written by the physical keyboard, but according to the virtual keyboard character. For example, if the first password character is 'x', the user will press 'k' on the physical keyboard as shown in Fig. 8.

4) The user will combine his real password with a camouflage character by using the activation key followed by a character equal to the value of the activation key. For example, if the activation key was 5 and the user password was xAvd4\$141 the user could enter 5t!C3wxAvd4\$141, xAvd4\$1415t!C3w or xAv5t!C3wd4\$141 as his password. The sequence 5t!C3w will be used as a camouflage for the password and can be added at any position to the real password.

5) After entering the password the decryption key will equal the number of entered alphabetic multiplied by the encryption key. For example, if the encryption key was 3 and the user entered 5 alphabetic the decryption key would be 15.

6) Finally, when the user submits the system will Decrypt the password by using Caesar decrypt equation as in Eq (2). Note that the password will be decrypted from right to left and the decryption key will be subtracted from the encryption key after decrypting each alphabetic. For example, if the decrypt key is 33 and the password is 'tupert24wertq', the letter 'q' will be decrypted with k equal 33 and 't' with k equal to 30. Then the system will omit the camouflaged character from the password to get the real password.

$$Dn(x) = (x-k) \text{ mod } 26 \quad (2)$$

As a result of this approach, if a person tries to comment shoulder surfing attack, he will not succeed; because the arrangements of the alphabet change after every click and the password is protected by the camouflaged character. Also, this technique is powerful against the keylogger that takes screenshots after each input.

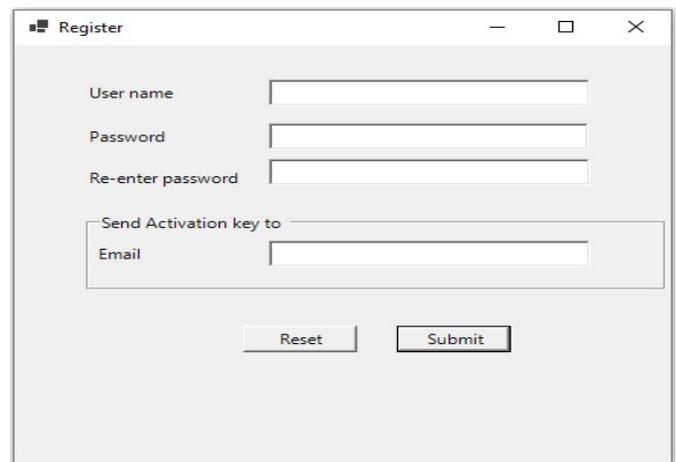


Fig. 5. System Register Screen.

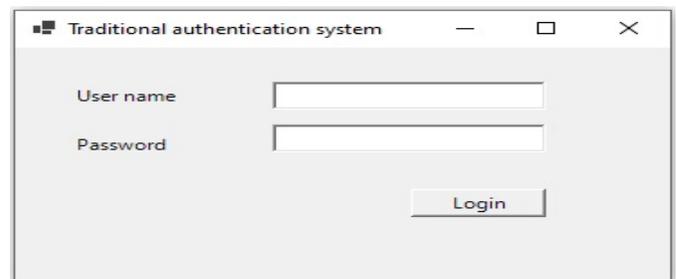


Fig. 6. Traditional System Login Screen.

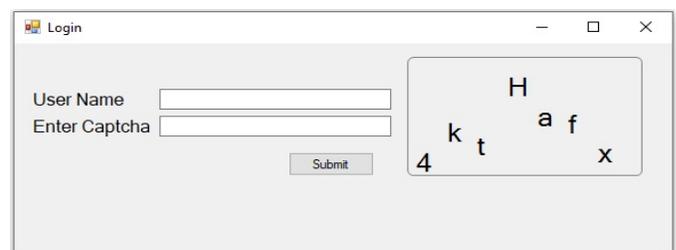


Fig. 7. Defensive Model Login Screen.

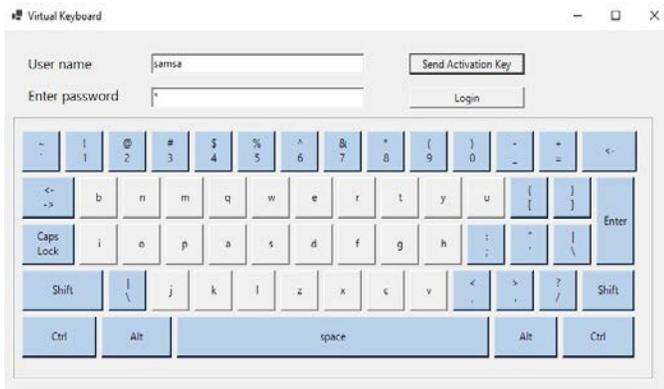


Fig. 8. The Virtual Keyboard when k=3.

## V. EXPERIMENTAL STUDY

The experimental study aims to evaluate the model against shoulder surfing attacks and to measure the usability of the proposed techniques. The following explains the experimental details:

1) Design two applications; the first, with a traditional login system, and the second, is based on the virtual keyboard; to be able to evaluate the proposed defensive techniques against shoulder surfing.

2) The experiment was done in a controlled environment in the lab and all participants entered 8-character passwords in both systems.

3) Experimental study with 65 participants divided into 20 PhD holders and 45 students. fifty-five of the participants will act as ordinary users, and 10 will act as shoulder surfing attackers. The experiment began with a brief explanation of the proposed defensive techniques. To become familiar with the system, the 55 participants are allowed to create their passwords and then login to the system approximately 3 times. Also, the 55 participants were informed that they are just allowed three attempts to login each time with each a count. If the three attempts failed, the account would be locked and the user has to try to login with a new account.

4) Measuring the success rate of the attack will clarify the models' susceptibility to shoulder surfing. To accomplish that, the 55 participants acted as victims, with 5 random students and 5 PhD holders acting as shoulder surfer attackers. The efficiency of the model is determined by analyzing and comparing the shoulder surfing success rate of the experiment with other studies.

## VI. RESULTS

To determine the model's usability two factors must be measured, entering time and login success rate. The success rate of the shoulder surfing attack will evaluate the model defensive technique. The result of these factors is used to compare four systems: The suggested traditional system which depends on static username and password memorized by the user, the defence proposed model and two models from other studies. For simplicity, the traditional login system will be represented with M1 and the defensive model with M2. While

the M1 password has a fixed number of characters which is 8, the M2 password is between 10 and 18 characters.

### A. Entering Time

The average time of the M1 model was 22.78, while M2 was 32.62 seconds as shown in Table I. The min and max time of the M1 system was 19.8 and 27.2 seconds, while in M2 was 30.1 and 36.6 seconds. The standard deviation (SD) in the entering time test of the two models was almost equal. By comparing the result it's clear that the M1 system is easier to use than the M2 model. Although the M1 test was better it has a fixed number of characters which is 8, while M2 has more characters even if the AK equals one. In general, the test result is affected by two factors:

- 1) The user typing speed.
- 2) The number of characters of the password.

### B. The Success Rate of Login

The login success rate is calculated by dividing the overall success login attempts over the total of all attempts of one participant. All 55 participants have given 3 attempts for login and they repeated this process for 5 different accounts. Table II shows the SD and mean values for M1 and M2 models. By comparing the mean value, M1 is better than M2 by 0.015. However, all participants have used the M1 model before in their life which makes it the most familiar model to them.

### C. Success Rate of Shoulder Surfing Attack

The experiment of shoulder-surfing attack was performed by 10 attackers and 55 users represent the victims. To keep the experiment real just 25 of the users were told that they will be watched to capture the login data. The attacker has to get the exact password and guessing is not allowed. The M1 model value was very high with 80.1%, while the defensive model M2 was 3.63% as shown in Table III.

The conducted result compared to the most related studies in terms of resisting shoulder surfing attacks and usability, support the M2 model technique in defending against shoulder surfing attacks as shown in Table IV.

TABLE I. THE RESULTS OF ENTERING TIME IN SECONDS

| Model | Min  | Max  | Mean  | SD   |
|-------|------|------|-------|------|
| M1    | 17.8 | 25.2 | 22.78 | 3.76 |
| M2    | 30.1 | 36.6 | 32.62 | 3.25 |

TABLE II. THE RESULTS OF THE LOGIN SUCCESS RATE

| Model | Mean  | SD   |
|-------|-------|------|
| M1    | 0.961 | 0.39 |
| M2    | 0.946 | 0.44 |

TABLE III. THE RESULTS OF THE SHOULDER-SURFING SUCCESS RATE

| Model | shoulder-surfing success rate |
|-------|-------------------------------|
| M1    | 80.1%                         |
| M2    | 3.63%                         |

TABLE IV. PROPOSED MODELS RESULT COMPARE TO OTHER STUDIES

| Study                  | Entry Time | Success Rate of login | The success rate of shoulder surfing |
|------------------------|------------|-----------------------|--------------------------------------|
| <b>Proposed models</b> |            |                       |                                      |
| M1                     | 22.78      | 96.1%                 | 80.1%                                |
| M2                     | 32.62      | 94.6%                 | 3.63%                                |
| <b>Other studies</b>   |            |                       |                                      |
| [28]                   | 23.2       | 64%                   | 26.4%                                |
| [29]                   | 3.66       | 97%                   | 16%                                  |

## VII. DISCUSSION

In this section, we analysed the data collected from the 65 participants, 55 acted as normal users and 10 as attackers. The M1 model has a faster entry time and a higher log in success rate than the M2 model; which makes the M1 a more friendly system. However, the M1 system is vulnerable to shoulder-surfing threats by 80.1% while M2 scores just 3.63%. The outcome obtained from the experiment indicates that using a visual keyboard with camouflage characters is a good defensive mechanism against the shoulder-surfing attack. It might be worth mentioning that the delay in M2 entry time is happening because the user must first check his email to get the AK key, then he must enter the password according to the dynamic virtual keyboard. This result is not enough to conclude that the M2 model is the best approach against shoulder-surfing, so it should be evaluated with other models [28][29] that are designed to fight against shoulder surfing. By comparing all models, it's obvious that the defence model M2 is the best defensive model with a success rate of shoulder surfing attack equal to 3.36%. However, it's the worst in entry time As shown in Fig. 9. Note that in the M2 model, the user must enter the password every time and can't use the autofill technique. Also, the model M2 aims to improve protection against shoulder surfing attacks, if it is done with human eyes or electronic devices like cameras. Furthermore, regardless of the defensive technique used users' behaviour is considered as the first line of defence against shoulder surfing attacks.

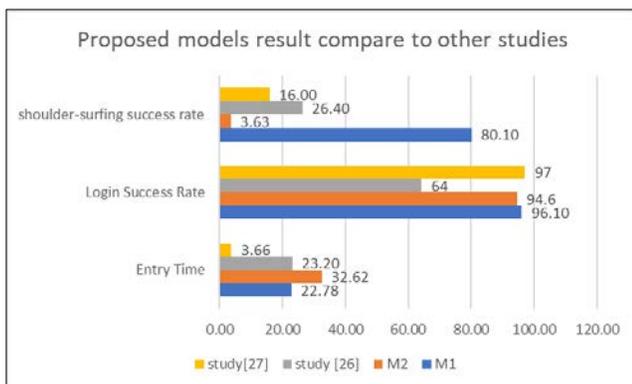


Fig. 9. The Proposed Model Result Compared to other Studies.

## VIII. CONCLUSION

In this paper, a new model that relies on camouflage characters and a virtual keyboard combined with the Caesar cipher is proposed as a defence method against shoulder

surfing attacks. An experiment with 65 users was conducted to evaluate the proposed defence model against a traditional login model. The evaluation depends on two factors: usability and the shoulder-surfing success rate. The results of the developed model were compared with two other studies that focus on shoulder surfing defence and use the same factors to measure their models. The obtained usability test indicates that the proposed defence model has the highest entry time compared to the other system, but it has the best result in preventing shoulder surfing. Depending on the results, the proposed defence model is recommended as the best solution for shoulder surfing attacks.

## REFERENCES

- [1] Dinei Florencio, and Cormac Herley, "A large-scale study of web password habits," Proceedings of the 16<sup>th</sup> international conference on World Wide Web, pp. 657-666, 2007.
- [2] J.Yan, A. Blackwell, R. Anderson and A. Grant, "Password memorability and security: empirical results," IEEE Security & Privacy, vol. 2, pp. 25-31, 2004.
- [3] Krishnapriya Kovalan et al., "A Systematic Literature Review of the Types of Authentication Safety Practices among Internet Users," International Journal of Advanced Computer Science and Applications, vol. 12, no. 7, 2021.
- [4] Eugene H.Spafford, "OPUS: Preventing weak password choices," Computers & Security, vol. 11, pp. 273-278, 1992.
- [5] Ari Kusyanti and Yustiyana April Lia Sari, "Creating and Protecting Password: A User Intention," International Journal of Advanced Computer Science and Applications, vol. 8, no. 8, 2017.
- [6] Arti Bhanushali, Bhavika Mange, Harshika Vyas, Hetal Bhanushali and Poonam Bhogle, "Comparison of Graphical Password Authentication Techniques," International Journal of Computer Applications, vol. 116, no. 1, 2015.
- [7] Manu Kumar, Tal Garfinkel, Dan Boneh and Terry Winograd, "Reducing shoulder-surfing by using gaze-based password entry," Proceedings of the 3<sup>rd</sup> symposium on Usable privacy and security, pp. 13-19, 2007.
- [8] Cheryl Hinds and Chinedu Ekwueme, "Increasing security and usability of computer systems with graphical passwords," Proceedings of the 45<sup>th</sup> annual southeast regional conference, pp. 529-530, 2007.
- [9] Huanyu Zhao and Xiaolin Li, "S3PAS: A Scalable Shoulder-Surfing Resistant Textual-Graphical Password Authentication Scheme," 21<sup>st</sup> International Conference on Advanced Information Networking and Applications Workshops, 2007.
- [10] Siddeeq Ameen Yousif and Laith Jasim Saud, "Computing Nodes and Links Appearances on Geodesics in Networks Topologies Using Graph Theory," Iraqi Journal of Computers Communications Control and Systems Engineering, vol. 12, no. 1, 2012.
- [11] Salim Istyaq and Khalid Saifullah, "A new hybrid graphical user authentication technique based on drag and drop method," International Journal of Innovative Research in Computer and Communication Engineering, vol. 6, 2016.
- [12] Suliman A. Alsuhibany, "Usability and shoulder surfing vulnerability of pattern passwords on mobile devices using camouflage patterns," Journal of Ambient Intelligence and Humanized Computing, pp. 1645-1655, 2020.
- [13] Jiya Gloria Kaka, Oyefolahan O. Ishaq and Joseph O. Ojeniyi, "Recognition-Based Graphical Password Algorithms: A Survey," IEEE 2nd International Conference on Cyberspac, 2021.
- [14] Jianwei Lai and Ernest Arko, "A Shoulder-Surfing Resistant Scheme Embedded in Traditional Passwords," Proceedings of the 54th Hawaii International Conference on System Sciences, p. 7144, 2021.
- [15] Aakansha S.Gokhale and Vijaya S.Waghmare, "The Shoulder Surfing Resistant Graphical Password Authentication Technique," Procedia Computer Science, vol. 79, pp. 490-498, 2016.
- [16] Dongmin Choi, Chang Choi and Xin Su, "Invisible Secure Keypad Solution Resilient Against Shoulder Surfing Attacks," 10<sup>th</sup> International

- Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, 2016.
- [17] Aravinda Thejas Chandra, G. Sneha, Srushti Anand and C. Yashaswini developed, "Real Time Eye Blink Password Authentication," International Journal of Research in Engineering Science and Management, vol. 4, no.7, 2021.
- [18] Anindya Maiti, Murtuza Jadliwala and Chase Weber," Preventing Shoulder Surfing using Randomized Augmented Reality Keyboards," IEEE International Conference on Pervasive Computing and Communications Workshops, 2017.
- [19] Eiji Hayashi, Rachna Dhamija, Nicolas Christin and Adrian Perrig," Use Your Illusion: Secure Authentication Usable Anywhere," Proceedings of the 4th symposium on Usable privacy and security, pp. 35-45, 2008.
- [20] Vishal Kolhe, Vipul Gunjal, Sayali Kalasakar and Pranjali Rathod, "Secure Authentication with 3D Password," International journal of Engineering Science and Innovative Technology, vol. 2, 2013.
- [21] Hung-Min Sun, Shiuan-Tung Chen, Jyh-Haw Yeh, and Chia-Yun Cheng, "A Shoulder Surfing Resistant Graphical Authentication System," IEEE Transactions on Dependable and Secure Computing, vol. 15, 2016.
- [22] Islam Abdalla, "Social Engineering Threat and Defense: A Literature Survey," Journal of Information Security, vol .9, 2018.
- [23] Fatima Salahdine and Naima Kaabouch, "Social Engineering Attacks: A Survey," Future Internet, 2019.
- [24] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann and Florian Alt, "Understanding Shoulder Surfing in the Wild: Stories from Users and Observers," Conference on Human Factors in Computing Systems, pp. 4254-4265, 2017.
- [25] Mohamed Khamis, Malin Eiband, Martin Zürn and Heinrich Hussmann, "EyeSpot: Leveraging Gaze to Protect Private Text Content on Mobile Devices from Shoulder Surfing," Multimodal Technologies and Interaction, 2018.
- [26] Arash Habibi Lashkari, Samaneh Farmand, Omar Bin Zakaria and Rosli Saleh, "Shoulder Surfing attack in graphical password authentication," International Journal of Computer Science and Information Security, vol. 6, no. 2, pp. 145-154, 2009.
- [27] Tonni Limbong and Parasian D.P. Silitonga, "Testing the Classic Caesar Cipher Cryptography using of Matlab," International Journal of Engineering Research & Technology, vol. 6, 2017.
- [28] Suliman A. Alsuhbany, "A Camouflage Text-Based Password Approach for Mobile Devices against Shoulder-Surfing Attack," Security and Communication Networks, 2021.
- [29] Emanuel von Zezschwitz, Alexander De Luca, Bruno Brunkow and Heinrich Hussmann, "Swipin: fast and secure pin-entry on smartphones," Proceedings of the 33<sup>rd</sup> Annual ACM Conference on Human Factors in Computing Systems, pp. 1403–1406, 2015.

# CovSeg-Unet: End-to-End Method-based Computer-Aided Decision Support System in Lung COVID-19 Detection on CT Images

Fatima Zahra EL BIACH, Imad IALA, Hicham LAANAYA, Khalid MINAOUI  
LRIT Associated Unit to the CNRST-URAC N29  
Faculty of Sciences, Mohammed V University in Rabat  
B.P.1014 RP, Rabat IT center  
Morocco

**Abstract**—COVID-19 epidemic continues to threaten public health with the appearance of new, more severe mutations, and given the delay in the vaccination process, the situation becomes more complex. Thus, the implementation of rapid solutions for the early detection of this virus is an immediate priority. To this end, we provide a deep learning method called CovSeg-Unet to diagnose COVID-19 from chest CT images. The CovSeg-Unet method consists in the first time of preprocessing the CT images to eliminate the noise and make all images in the same standard. Then, CovSeg-Unet uses an end-to-end architecture to form the network. Since CT images are not balanced, we propose a loss function to balance the pixel distribution of infected/uninfected regions. CovSeg-Unet achieved high performances in localizing COVID-19 lung infections compared to others methods. We performed qualitative and quantitative assessments on two public datasets (Dataset-1 and Dataset-2) annotated by expert radiologists. The experimental results prove that our method is a real solution that can better help in the COVID-19 diagnosis process.

**Keywords**—Deep learning; COVID-19; loss function; balanced data

## I. INTRODUCTION

In December 2019, a viral pneumonia epidemic of unknown etiology emerged in Wuhan city, Hubei province, China [1]. On January 9, 2020, the World Health Organization (WHO) and Chinese Health Authorities officially announced the discovery of a new coronavirus. This pneumonia is an infectious disease caused by a virus identified under the name SARS-CoV-2 (Severe Acute Respiratory Syndrome CoronaVirus-2) by the ICTV (International Committee on Taxonomy of Viruses) [2], and causing a disease called COVID-19 (COroNaVirus Disease 2019). SARS-CoV-2 belongs to the coronavirus family. The reservoir of this virus is probably animal. Although SARS-CoV-2 closely resembles a virus detected in a bat, the animal that transmits it to humans has yet to be identified with certainty. Several research studies suggest that the pangolin, a small mammal eaten in southern China, could be involved as an intermediate host between bats and humans.

The new coronavirus has been confirmed to be transmitted between humans [3], and this is done mainly by air or by close

contact with a contagious subject. Smaller particles can also be emitted in the form of aerosols during speech or during coughing efforts, which would explain that the virus could persist suspended in the air in an unventilated room. Finally, the virus can retain infectivity for a few hours on inert surfaces from where it can be transported by the hands. According to data from the World Health Organization, updated up to 24 hours on June 18, 2021, COVID-19 has affected 220 countries and territories, causing 178,584,744 people to be infected and 3,866,607 deaths worldwide. The overall number of people recovered is 163,102,134. Currently, the active cases are 11,616,003 of which 99.3% in mild condition and 0.7% in serious or critical condition, which poses a great threat to international human health.

Due to the vaccination process slowness, the high rate of virus contamination, and the appearance of new dangerous COVID-19 mutations, it is essential to detect and identify the disease at an early stage so that suspected patients do not infect the healthy population. As a result, new requirements for the prevention and control strategy must be put in place. Reverse Transcription Polymerase Chain Reaction (RT-PCR), gene sequencing for respiratory, or blood samples confirm the diagnosis of COVID-19. However, the false negatives of the RT-PCR [4], the delay in obtaining the results, and the tests carried out on people not strongly suspected of being infected with COVID-19 imply that numerous COVID-19 patients would not be identified quickly to isolate them from others. In addition, given the rapid and contagious spread of the virus, they present a real threat to infect a larger population, especially in areas with high epidemics. On the other hand, chest examinations quickly established themselves as an interesting diagnostic tool, given the characteristic presentation of COVID-19 lesions [5]. These tests can identify lesions, underlying conditions and complications associated with acute airway conditions. Consequently, the use of CT in particular High Resolution Computed Tomography (HRCT) could provide enormous help to radiologists [7] for the diagnosis, follow-up or investigation of pulmonary complications in patients suspected or confirmed of COVID-19. Thus, the development of an artificial intelligence (AI) method based on deep learning could help them enormously to assess the degree of lung damage caused by COVID-19.

According to [6], the authors indicated that CT images can be used to detect COVID-19 even before certain clinical symptoms are observed. Typical signs of COVID-19 appear in CT images as unilateral, multifocal, and terminal Ground Glass Opacity (GGO). This is a hazy cloud above the lungs that indicates a variety of problems, and may mean that the lungs are partially filled with inflamed material, and there is thickening in lung tissue or partial breakdown of the alveoli and tiny air sacs of the lungs. Pleural effusion, lymphadenopathy, and condensation [3], which are air spaces in the lungs filled with a substance, usually pus, blood or water, surrounded by an opaque edge of frosted glass, and although this is a common feature of lung disease, it may be more characteristic of COVID-19. To detect COVID-19 disease at an early stage, it is necessary to detect and locate these pathological changes in a short time. The growing number of patients and the limited number of well-trained expert radiologists in most hospitals prevent and slow down the process of early detection. Indeed, the use of deep learning methods for the automatic segmentation of the COVID-19 CT model has become paramount, and may offer an effective solution to identify and locate signs of COVID-19 in CT images [8].

In this paper, a new efficient method of COVID-19 diagnostic using Deep Learning network is proposed. Section II presents the related works, Section III shows problem statement, and Section IV explains in detail the proposed method. Section V describes simulations experiments. Section VI discusses the obtained results. Finally, the summary and future works are delineated in Section VII.

## II. RELATED WORK

In the literature, numerous methods of segmentation based on deep learning networks have been used to process and analyze chest X-ray or CT images for the COVID-19 diagnosis [9]. These methods mainly consist of delineating the regions of interest in these images, such as lobes, bronchopulmonary segments, lung, and infected regions or lesions for further quantification and evaluation.

CT provides detailed and high definition three dimensional images to detect COVID-19. Among the segmentation methods used for the diagnosis of COVID-19 we cite, U-Net [10], UNet++ [11], V-Net [12]. The authors of [9] have proposed a 3D architecture of U-Net using inter-slice information; this method consists in replacing the conventional layers of U-Net by a 3D version. In [12], the authors proposed the V-Net architecture, in which they used the residual blocks as the basic convolutional block, and optimized the network by a loss of dice. In [13], the authors proposed the Attention U-Net method, which captures fine structures to locate lesions and pulmonary nodules in medical images. Generally, the large number of well-labeled images is the key to forming an efficient and robust segmentation network. In the case of COVID-19 image segmentation, the data used during the training phase is limited and often unavailable because manual lesion delineation is a difficult operation and requires a lot of time.

Several other research works obtaining reasonable segmentation results have been proposed in this context. The lung segmentation field is experiencing a lack of labeled medical images, as a result, semi supervised and unsupervised methods are very favorable and recommended in studies on COVID-19, as in [10], the authors used an unsupervised method to generate pseudo-segmentation masks for the images. In [14], the authors proposed a new COVID-19 lung infection segmentation network called Inf-Net to detect infected regions from chest CT images. This method uses a parallel decoder for aggregating high-level features and generating a global map. Then, it uses a semi-supervised segmentation framework based on a propagation strategy chosen at random to overcome the lack of labeled data. In [15], the authors proposed a computer-assisted diagnosis (CAD) system based on the YOLO predictor to detect and diagnose COVID-19. The CAD method calls the data balancing regularizations, transfer learning, and augmentation to improve the overall diagnostic performance for COVID-19. The authors of [16] proposed a synergistic approach based on deep meta-learning to accelerate the detection of COVID-19 cases. This approach uses contrastive learning with a pre-trained ConvNet encoder for the classification of COVID-19 cases. In [17], the authors proposed a computer-aided detection (CAD) method to assist radiologists to automatically detect COVID-19 on the chest X-ray images. The proposed method uses the DLs: the Discrimination-DL to extract lung features from chest X-ray images, and the Localization-DL to localize and assign the infected lung region. In [18], the authors built prognosis models to predict the patients' severity outcomes. The proposed method is based on deep learning in the CT image segmentation process for COVID-19 pneumonia, and it uses datasets from multiple institutions worldwide to validate the proposed models. In [19], the authors propose a CovFrameNet framework to detect COVID-19 cases using CT images, which incorporates an image preprocessing mechanism and a deep learning model for smoothing, denoising, feature extraction, classification, and performance measurement.

## III. PROBLEM STATEMENT

Our main goal through this paper is to diagnose COVID-19 lung infection in chest CT images. In this regard, we use an architecture similar to that of the U-Net [10] approach called CovSeg-Unet method, U-Net is considered as the most commonly used algorithm in medical image segmentation. U-Net is a symmetrical encoder/decoder architecture consisting of several stages. Each stage of the encoder performs a set of operations such as, convolution, normalization, max pooling, activation, concatenation etc. In parallel, each stage of the decoder performs deconvolution operations. U-Net method uses jump connections allowing the exploitation of local and global information. These connections concatenate the subsampling characteristics of the contraction path with those of the up-sampling of the expanding path.

The general idea of the CovSeg-Unet approach is described as follows. Let  $S$  be a learning space which contains a set of  $n$  images  $X = X_1 \dots X_n$ , and  $n$  corresponding ground-truth masks  $Y = Y_1 \dots Y_n$ . From the  $Y$  ground-truth masks, the network learns the lung infections distribution of the  $X$

learning images to establish an image-to-image mapping relationship between  $X$  and  $Y$ , this map is defined as follows:  $\Phi = f_{dec} \circ f_{enc}$ .

$f_{enc}$  is the encoder function that learns the characteristic vectors of infected lung regions to establish a functional space. While  $f_{dec}$  is the decoder function, which learns the spatial localization of features to better locate the infected/uninfected region. And  $\Phi$  is the set of probabilities extracted from an input biomedical image, and is represented by  $\Phi(f) = P_1, P_2, \dots, P_n$ , where  $P_i = \{0,1\}$  are the probabilities given to the last convolutional layer of classification by the nonlinear function SoftMax. The key idea here focuses on the spatiotemporal modeling of characteristic points that can represent lung infections.

#### IV. PROPOSED METHOD

In this section, we present the CovSeg-UNet approach that we propose to diagnose COVID-19 from CT images. The CovSeg-UNet approach architecture is shown in Fig. 1. Generally, the CovSeg-UNet approach architecture's is made up of two blocks; the first one consists of preprocessing the CT images. While the second one performs all encoder/decoder operations to learn high level features from training sets, and locate the spatial information. The CovSeg-UNet network details are shown.

##### A. Preprocessing

One of the major problems encountered in the deep learning is processing of biomedical images that coming from several machines with different acquisition parameters. To deal with this problem, we apply a preprocessing step on these images to improve the learning of the COVID-19 suspect recognition model despite of the data heterogeneity. This step of preprocessing consists of two steps; the first one concerns the normalization of the signal that processes the intensity of each scanner CT voxel. In this step, we chose the pulmonary window value from Table I to separate the lungs from the other organs. Since each CT scanner has its own Hounsfield (HU) units, consequently, the data collected in different hospitals will have different HUs. For this reason, we are

using a multi-valued window, ( $W_L$  is the window center value, and  $W_W$  is the window width), the  $W_L$  value is randomly assigned from -600 to -500, while the  $W_W$  value is fixed at 1200. In the second preprocess step, we normalizing the CT images to be in [0, 255]. Through this process we normalize multiple images of different scanners to the same standard, by separating the lungs from other organs, removing unnecessary information (CT features, etc.) or/and noise, increasing images, and improved accuracy.

##### B. CovSeg-UNet Architecture

In order to detect COVID-19 lung infections using CT images, we propose the CovSeg-UNet approach, which characterized by an end-to-end architecture based on one of the most robust approaches in biomedical image segmentation that is U-Net. The network of the CovSeg-UNet approach is supplied as input by pre-processed CT images and their ground-truth masks. Our proposed method relies on an encoder to extract contextual feature maps from pre-processed images to reduce the dimensions of CT images, and a decoder to locate feature map information in the image. Table II shows the detailed architecture of our encoder/decoder.

Generally, the encoder is made up of four residual blocks (ResBlock), already pre-formed on the ImageNet database, the use of these blocks allows us to avoid the disappearance/explosion gradient, to preserve the local information, to improve precision by increasing the depth of the network, and to optimize the formation of layers. Each residual block is made up of two blocks (conv\_bloc and identity\_bloc), and is expressed by  $R(I) + I$ , where  $I$  is the input vector, and  $R(\cdot)$  represents the mapping from input  $I$  to the output of the residual unit. However, the decoder has two inputs, one input from the parallel layer of the encoder and a second from the previous layer of the decoder. Finally, the decoder output is sent to the SoftMax activation function (see equation (1)) for the prediction of the region infected with COVID-19. Algorithm 1 describes the training steps of the proposed model.

$$\sigma(\vec{z})_i = \frac{e^{x_i}}{\sum_{j=1}^k e^{z_j}} \quad (1)$$

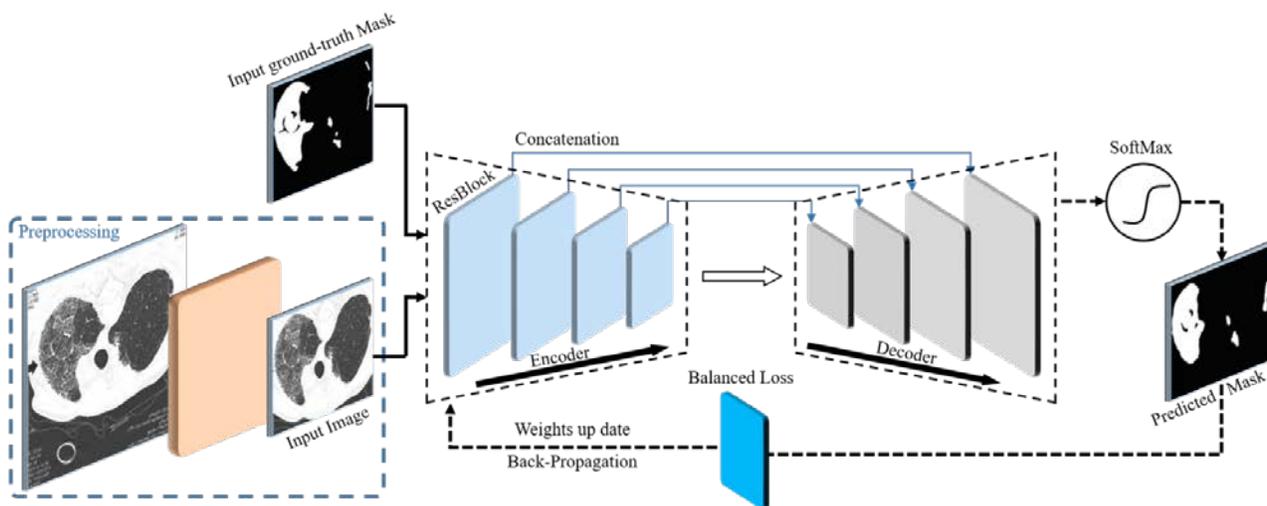


Fig. 1. Overview of the Proposed Framework for Diagnosing COVID-19 from CT Images.

TABLE I. DIFFERENT HOUNSFIELD VALUES OF DIFFERENT SUBSTANCES

| Substance                | Hounsfield Unit (HU) |
|--------------------------|----------------------|
| Air                      | -1000                |
| Bone                     | +700 to +3000        |
| Lungs                    | -500                 |
| Water                    | 0                    |
| Kidney                   | 30                   |
| Blood                    | +30 to +45           |
| Grey matter              | +37 to +45           |
| liver                    | +40 to +60           |
| White matter             | +20 to +30           |
| Muscle                   | +10 to +40           |
| Soft Tissue              | +100 to +300         |
| Fat                      | -100 to -50          |
| Cerebrospinal fluid(csf) | 15                   |

TABLE II. DETAILED ARCHITECTURE OF THE PROPOSED METHOD. WHERE THE ENCODER IS THE RESIDUELLE BLOCK SIMILAIRE TO RESNET50[20] ARCHITECTURE, AND THE DECODER BLOCK IS A SUCCESSION OF BATCHNORMALISATION, RELU, CONV2D, BATCHNORMALISATION, AND RELU OPERATIONS

| Stages  | Layers                                                                                                                                                       | Output size                                                             |
|---------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
| Initial | Conv2d(7 x 7), stride = 2<br>BatchNormalisation<br>ReLu<br>MaxPoling                                                                                         | (64 x 64 x 64)                                                          |
| Encoder | Stage 1 : x2 ResBlock(64 x 64 x 256)<br>Stage 2 : x3 ResBlock(128 x128 x512)<br>Stage 3 : x5 ResBlock(256x256 x1024)<br>Stage 4 : x2 ResBlock(512x512 x2048) | (64 x 64 x 256)<br>(32 x 32 x 512)<br>(16 x16 x 1024)<br>(8 x 8 x 2048) |
| Decoder | Stage 1 : UpSampling(16 x 16 x 256)<br>Stage 2 : UpSampling(32 x 32 x 128)<br>Stage 3 : UpSampling(64 x 64 x 64)<br>Stage 4 : UpSampling(128 x 128 x 32)     | (32 x 32 x128)<br>(64 x 64 x 64)<br>(128 x128 x32)<br>(256 x 256 x16)   |
| Final   | Conv2d(256 x 256 x 16)<br>Softmax Activation(256 x 256 x 1)                                                                                                  | (256 x 256 x 1)<br>(256 x 256 x 1)                                      |

Where  $\vec{z}$  is the input vector of the softmax function ( $z_0, \dots, z_k$ ), all  $z_i$  values can take any real value.  $e^{z_i}$  is applied to have a positive value for each element of the input vector. The term at the formula bottom is the normalization term, which makes it possible to have a  $Sum = 1$  of all the output values (are each in the range (0, 1)) of the function, thus constituting a valid probability distribution,  $k$  represents the number of classes in the multiclass classifier.

**Algorithm 1:** Training procedure for CovSeg-UNet method

```

input : X : Image CT; Y : Label; N : Batch-size; λ_1 ; λ_2 ;
 Lr : Learning-rate; W_0 : Initial weights;
output: P_L : Predicted mask;
1 begin
2 $(X_{train}, Y_{train}), (X_{valid}, Y_{valid}) \leftarrow \text{split}((X, Y),$
 $\text{split-size}=0.2)$
3 while epoch ≤ 200 do
4 for mini batch sample $x_k \{x_{train}, x_{valid}\}_{k=1}^N$ do
5 $z_1 = f_{encoder}(x_k)$
6 $z_2 = f_{decoder}(z_1)$
7 predict(z_2) with SoftMax equation
8 Compute Δ_w the stochastic gradient by
 minimizing the loss function eq. 2
9 Update weights
10 $w \leftarrow w + \alpha \cdot \text{AdamOptimiser}(w, \Delta_w)$

```

C. Loss

The imbalanced class problem is considered as the major challenge in the detection process of lung infections because the distribution of infected/uninfected regions is highly skewed (the infected regions vary between 0% and 20% of the pixels of the lung image). So, if the loss function does not consider this problem, the model will classify the majority of pixels as uninfected regions, and become overfit. For this reason, we use a class-balanced cross-loss function and the penalty factor  $\lambda$ .

The loss function is defined as a weighted sum of two loss functions; the balanced cross loss  $L_{BCE}$  and the inverse cross loss  $L_{ICE}$ :

$$L_B = \lambda_1 \cdot L_{BCE} + \lambda_2 \cdot L_{ICE} \quad (2)$$

Respecting to the cross-loss, we use the balanced cross-entropy to overcome noise in biomedical images, we also use the balance parameter  $W$  to balance the pixel distribution of infected/uninfected regions in  $L_{BCE}$ :

$$L_{BCE} = -w_c \sum_{i \in y_c} q(y_i = 1 | s) \log(p(y_i = 1 | s)) - w_{cn} \sum_{i \in y_{nc}} q(y_i = 0 | s) \log(p(y_i = 0 | s)) \quad (3)$$

Where  $q(y_i = i | s)$  is the ground-truth mask of the sample  $S$ .  $Y_i = 0, 1$ ,  $p(y | s)$  is the probability map produced by the function softmax,  $w$  represents the balancing parameter  $w(k) = S/KS(k)$ .  $S$  is the samples number in the training set, and  $S(K)$  represents the samples number in the class  $K$ . However, cross-entropy relies heavily on the accuracy of the annotation. When the data is mislabeled,  $q(k | s)$  will not be able to represent the true class distribution, which will lead the cross-entropy  $p(k | x)$  to learn this incorrect distribution type. To deal with this problem, we use reverse learning to know which classes the input  $x$  does not belong to. Inverse cross-entropy is defined as follows:

$$L_{ICE} = \sum_{i=1}^k p(y_i = 1 | s) \log q(y_i = 1 | s) \quad (4)$$

The weighted combination of entropies ( $L_{ICE}$  and  $L_{BCE}$ ) in the loss function allowed a good convergence of the gradient and relevant learning. The  $L_{BCE}$  term efficiency exhibits in the distribution balance of infected and uninfected classes, while the  $L_{ICE}$  term strength appears in the resistance against noise caused by scanner settings. As a result, the balanced symmetric entropy function obtained high performance in the COVID-19 lung infections localization on the test data set.

V. EXPERIMENT

In this section, we explain the different experiments carried out to evaluate the performance of our proposed approach under different simulation scenarios. We start by describing the used dataset and experimental environment of the simulation. Subsequently, we perform quantitative study of the hyper-parameters to show their effects on the model learning. Finally, we present the different performance metrics used to assess the efficiency and robustness of our approach.

A. Datasets

In this work, we have applied our method on two Datasets.

Dataset-1: this dataset contains two versions [21]. The first version is published on April 2, 2020, comprising 100 CT images (of 40 COVID-19 patients) of size  $512 \times 512$  labeled by radiologists, these radiologists have defined three tags: pleural effusion in frosted glass (= 3), consolidation (= 2), and mask value (= 1). The second version is released on April 14, 2020, comprising 829 CT images (of 9 COVID-19 patients) sized  $630 \times 630$ . Radiologists have tagged 373 images with COVID-19 pulmonary symptoms and the rest of the images as normal cases.

Dataset-2: this dataset is publicly available, contains 20 CT volumes with more than 1800 slices collected from 40 different COVID patients [22]. Each CT slice is of size  $512 \times 512$  labeled by expert radiologists to mark regions for infections.

The first column images of Fig. 2 represent the original images, while the second column images represent their corresponding ground-truth masks. Two examples of images of normal people are shown in the third and fourth rows. The first and second rows include two COVID-19 images, where the COVID-19 regions are the white and gray regions of the ground-truth masks, while the healthy regions are the black pixels (note that if a person is in good health his ground-truth mask will be completely black).

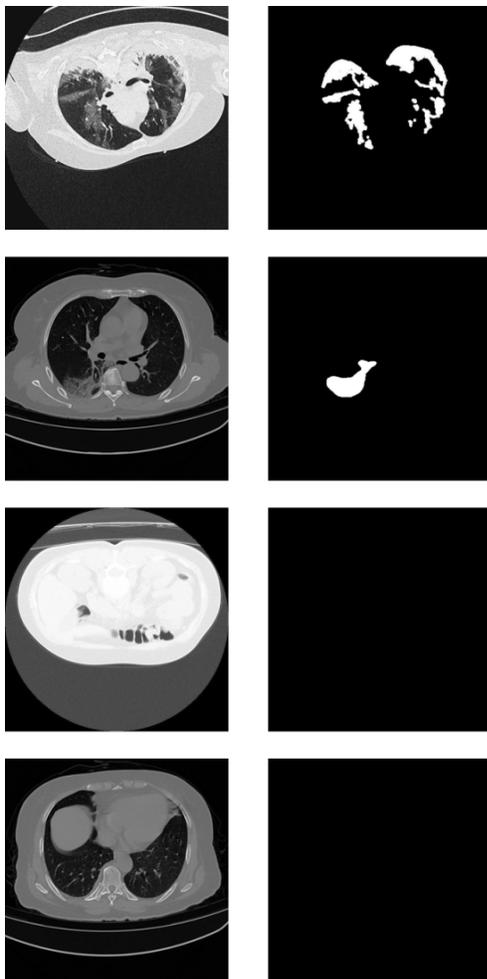


Fig. 2. Example of Images belong to Dataset-1 and Dataset-2.

In our simulations, we merged the two versions of dataset-1 to form a new dataset, while keeping 60% of these images for the training, 20% for the validation, and 20% for the test. We implemented our method in the Keras simulation environment with the TensorFlow back-end using the Python 2.7 programming language. Simulations were run on an infrastructure equipped with a Tesla P-100 GPU card, and 16 GB RAM memory.

*B. Hyper-parameters Setting*

In deep learning, hyper-parameters of the deep neural network crucially influence the performance of the network. In this part, we carried out several experiments to choose the best values of the Hyper-parameters allowing improving the performances of our method. In this regard, we fixed the number of epochs and the batch size respectively at 100 and 64. We simulated and compared the different precision obtained values with different optimizers, different learning rates, and different values of  $\lambda_1, \lambda_2$  of the  $L_B$  loss function.

TABLE III. THE CHOICE OF THE BEST COMBINATION FOR INPUT HYPERPARAMETERS, BEST ACCURACY SHOWN IN BOLD FONT

| Optimizer               | Learning-rate                                      | Loss (LB)                                                                                                                        | Accuracy                          |
|-------------------------|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|-----------------------------------|
| ADAM<br>SGD<br>Adadelta | Lr = 0.0001                                        | $\lambda_1 = 0.5, \lambda_2 = 0.5$                                                                                               | <b>0.9445</b><br>0.8921<br>0.9032 |
| ADAM                    | Lr = 0.001<br>Lr = 0.00001<br><b>Lr = 0.000001</b> | $\lambda_1 = 0.5, \lambda_2 = 0.5$                                                                                               | 0.944<br>0.924<br><b>0.964</b>    |
| ADAM                    | Lr=0.000001                                        | $\lambda_1 = 0.4, \lambda_2 = 0.2$<br>$\lambda_1 = 0.5, \lambda_2 = 0.5$<br><b><math>\lambda_1 = 0.3, \lambda_2 = 0.5</math></b> | 0.957<br>0.964<br><b>0.991</b>    |

From Table III, we observe that the values of the hyper-parameters  $\lambda_1 = 0.3, \lambda_2 = 0.5$ , and the use of the ADAM [23] optimizer with a learning rate equal to 0.000001 showed the best performance in term of accuracy.

*C. Performance Metrics Evaluation*

To assess the efficiency and robustness of our proposed approach, we use the following performance metrics: Accuracy [24], Sensitivity [25], Matthews Correlation Coefficient [25], and Dice [26]. By definition, higher values of these metrics imply a better segmentation quality. The mathematical formulas of these metrics are respectively expressed below:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{5}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{6}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

$$Dice = \frac{2 * TP}{2 * (TP + FN + FP)} \tag{8}$$

VI. RESULTS

In this section, we evaluate the effectiveness of the proposed method by performing qualitative and quantitative studies on an open-source benchmark. The obtained results of our method are compared with those of existing methods in the state of the art.

A. Ablation Study

To show the importance of each component of our approach, we did an ablation study on the dataset-1 by evaluating the following performance metrics; Accuracy, Dice, Sensitivity, and Precision. The study of ablations was subdivided into three possible cases. In the first case, we added the preprocessing block without using the loss function ( $L_B$ ) during the learning phase, whereas in the second case, in the learning phase we introduced the loss function without using the preprocessing block. In the latter case, we used the preprocessing block and the loss function ( $L_B$ ) (see Fig. 3). From these simulation cases, we notice that the data preprocessing step has a remarkable effect on the model performances. The first case shows the overfitting phenomenon that occurred when it exceeds epoch 40, and which led to a degradation in the performance of the model (see Fig. 4). The results of this study are illustrated in Table IV.

On the other hand, the result of the third case shows that the use of the loss function with the preprocessing block helps to avoid the overfitting problem, and consequently, to improve the performances of the model and to have good results. Fig. 5 and Fig. 6 show the qualitative results of our method on different test samples, the first column (1) represents the lung images, the second column (2) shows the ground-truth mask, and the last column (3) corresponds to the predicted lung infection mask of COVID-19. These qualitative results prove the robustness and the efficiency of our diagnostic method of the region infected by COVID-19.

TABLE IV. ABLATION STUDY ON THE DATASET-1

| Cases                                  | Accuracy     | Dice         | MCC          | Sensitivity  |
|----------------------------------------|--------------|--------------|--------------|--------------|
| Preprocessing, BinaryCrossEntropy      | 0.890        | 0.514        | 0.432        | 0.741        |
| $L_B$ , without Preprocessing,         | 0.934        | 0.632        | 0.590        | 0.842        |
| <b>Preprocessing, <math>L_B</math></b> | <b>0.991</b> | <b>0.833</b> | <b>0.851</b> | <b>0.982</b> |

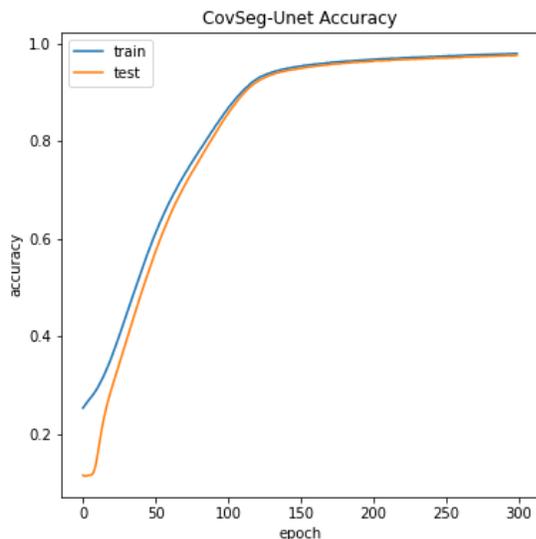


Fig. 3. Accuracy of Simulation with Data Preprocessing Step.

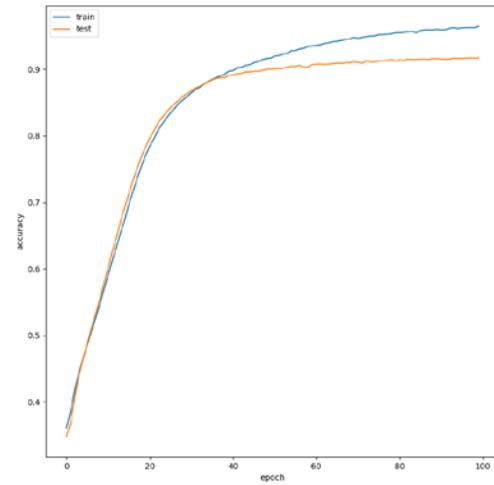


Fig. 4. Accuracy of Simulation without Data Preprocessing Step.

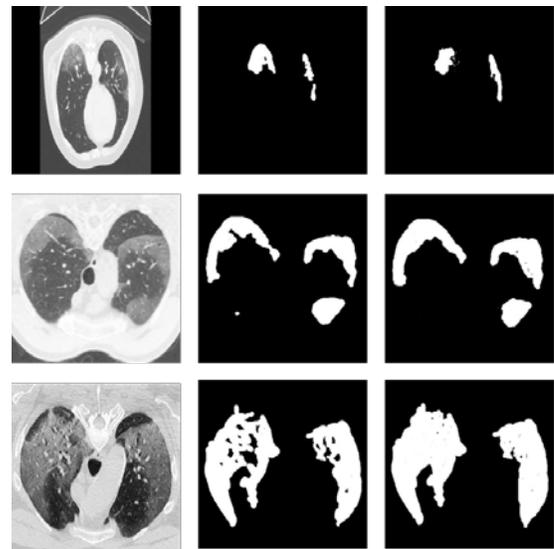


Fig. 5. Results Examples obtained by CovSeg-Unet on Dataset-1.

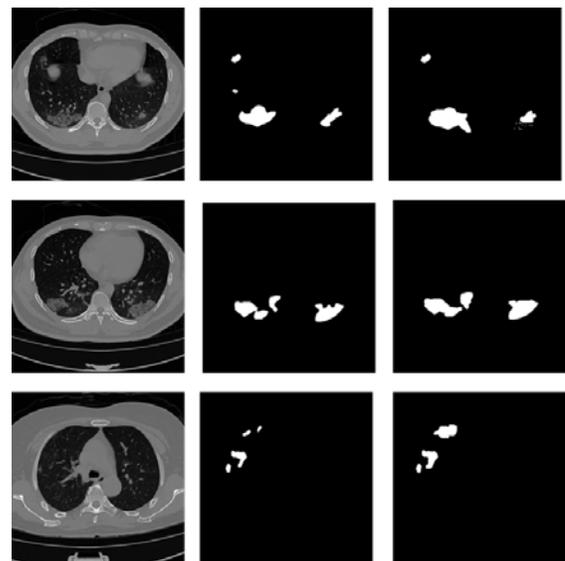


Fig. 6. Results Examples obtained by CovSeg-Unet on Dataset-2.

### B. Comparison with Baseline Methods

In this part, we compare our segmentation method of CT images with the reference segmentation methods such as U-Net basic [10], DenseUNet [27], Attention U-Net [13], and UNet++ [11]. Table V represents obtained results by the different methods on dataset-1 and dataset-2. As results, the proposed method outperforms other methods in terms of performance measures. Dice, Sensitivity, and Accuracy metrics reach 83.3%, 98.2% and 99.1% respectively on dataset-1, and 83.4% 89.1% 98.3% on dataset-2.

TABLE V. PERFORMANCES COMPARISON AGAINST BASELINE ARCHITECTURES ON DATASET-1 AND DATASET-2

|           | Methods              | Dice         | Sensitivity  | Accuracy     |
|-----------|----------------------|--------------|--------------|--------------|
| Dataset-1 | U-Net [10]           | 0.708        | 0.678        | 0.865        |
|           | DenseUNet [27]       | 0.660        | 0.607        | 0.651        |
|           | Attention U-Net [13] | 0.560        | 0.623        | 0.632        |
|           | UNet++ [11]          | 0.815        | 0.857        | 0.903        |
|           | <b>CovSeg-Unet</b>   | <b>0.833</b> | <b>0.982</b> | <b>0.991</b> |
| Dataset-2 | UNet [10]            | 0.712        | 0.665        | 0.747        |
|           | DenseUNet [27]       | 0.610        | 0.607        | 0.715        |
|           | Attention U-Net [13] | 0.631        | 0.723        | 0.890        |
|           | UNet++ [11]          | 0.815        | 0.887        | 0.968        |
|           | <b>CovSeg-Unet</b>   | <b>0.834</b> | <b>0.891</b> | <b>0.983</b> |

### C. Comparison with other Methods

Many studies have been done to diagnose COVID-19. To prove the robustness of our method, we carried out a comparative study with different approaches such as Inf-Net [14] and Automatic [17]. We simulated these approaches using their open-source implementation. The quantitative results obtained by the different methods are shown in Table VI. Dice, Sensitivity, and Accuracy are the performance metrics to be evaluated in this benchmarking study. From Table VI we notice that the CovSeg-Unet method reaches an Accuracy = 0.991 on Datasets-1, and an Accuracy = 0.983 on Datasets-2. The obtained values in these simulations prove that the CovSeg-Unet method outperforms other approaches in terms of Dice, Sensitivity, and Accuracy.

TABLE VI. PERFORMANCES COMPARISON AGAINST EXISTING APPROACHES ON DATASETS-2.

| Methods            | Dice         | Sensitivity  | Accuracy     |
|--------------------|--------------|--------------|--------------|
| Inf-Net [14]       | 0.579        | 0.877        | —            |
| Automatic [17]     | 0.714        | 0.733        | 0.739        |
| <b>CovSeg-Unet</b> | <b>0.834</b> | <b>0.891</b> | <b>0.983</b> |

## VII. CONCLUSION

In this work, we address a more difficult task in the segmentation of limited and unbalanced biomedical images. To cope with this task, we have proposed an end-to-end architecture similar to U-Net, the proposed method network learns the discriminating features of lung infections from CT images to establish an image-to-image mapping relationship. We used the ResNet50 architecture to preserve local information and avoid the issue of fading gradients. To improve the learning of the discriminating features of the network, we introduced a preprocessing block to remove noise and unnecessary information which blows up the performance of the network. In order to strengthen the model to be learned

from non-equilibrium data, we have proposed a loss function  $L_B$ . Experimental results on two datasets demonstrated the effectiveness of the CovSeg-Unet method in locating COVID-19 infected regions. The quantitative and qualitative results obtained by comparing CovSeg-Unet method with the others methods prove the efficiency of our method, which can be a real solution to detect, diagnose and locate regions infected with COVID-19.

### REFERENCES

- [1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu et al., "Clinical features of patients infected with 2019 novel coronavirus in wuhan, china," *The lancet*, vol. 395, no. 10223, pp. 497-506, 2020.
- [2] N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu et al., "A novel coronavirus from patients with pneumonia in china, 2019," *New England journal of medicine*, 2020.
- [3] J. T. Wu, K. Leung, and G. M. Leung, "Nowcasting and forecasting the potential domestic and international spread of the 2019-ncov outbreak originating in wuhan, china: a modelling study," *The Lancet*, vol. 395, no. 10225, pp. 689-697, 2020.
- [4] T. Liang, H. Cai, Y. Chen, Z. Chen, Q. Fang, W. Han et al., "Handbook of covid-19 prevention and treatment. 2020," *The First Affiliated Hospital, Zhejiang University School of Medicine. Compiled According to Clinical Experience*, 2020.
- [5] F. Pan, T. Ye, P. Sun et al., "Time course of lung changes on chest ct during recovery from 2019 novel coronavirus (covid-19) pneumonia [e-pub ahead of print], *radiology*," 2020.
- [6] S. Salehi, A. Abedi, S. Balakrishnan, and A. Gholamrezanezhad, "Coronavirus disease 2019 (covid-19): a systematic review of imaging findings in 919 patients," *American Journal of Roentgenology*, vol. 215, no. 1, pp. 87-93, 2020.
- [7] J. P. Kanne, "Chest ct findings in 2019 novel coronavirus (2019-ncov) infections from wuhan, china: key points for the radiologist," 2020.
- [8] F. Shan, Y. Gao, J. Wang, W. Shi, N. Shi, M. Han, Z. Xue, D. Shen, and Y. Shi, "Lung infection quantification of covid-19 in ct images with deep learning," *arXiv preprint arXiv:2003.04655*, 2020.
- [9] Ö. Çiçek, Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424-432.
- [10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234-241.
- [11] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3-11.
- [12] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565-571.
- [13] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz et al., "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.
- [14] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2626-2637, 2020.
- [15] M. A. Al-antari, C.-H. Hua, J. Bang, and S. Lee, "Fast deep learning computer-aided diagnosis of covid-19 based on digital chest x-ray images," *Applied Intelligence*, vol. 51, no. 5, pp. 2890-2907, 2021.
- [16] M. Shorfuzzaman and M. S. Hossain, "Metacovid: A siamese neural network framework with contrastive loss for n-shot diagnosis of covid-19 patients," *Pattern recognition*, vol. 113, p. 107700, 2021.

- [17] Z. Wang, Y. Xiao, Y. Li, J. Zhang, F. Lu, M. Hou, and X. Liu, "Automatically discriminating and localizing covid-19 from community acquired pneumonia on chest x-rays," *Pattern recognition*, vol. 110, p.107613, 2021.
- [18] K. Gong, D. Wu, C. D. Arru, F. Homayounieh, N. Neumark, J. Guan, V. Buch, K. Kim, B. C. Bizzo, H. Ren et al., "A multi-center study of covid-19 patient prognosis using deep learning-based ct image analysis and electronic health records," *European journal of radiology*, vol. 139, p. 109583, 2021.
- [19] O. N. Oyelade, A. E. Ezugwu, and H. Chiroma, "Covframenet: An enhanced deep learning framework for covid-19 detection," *IEEE Access*, 2021.
- [20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.
- [21] "covid-19 ct segmentation dataset. available: <https://medical-segmentation.com/covid19/>,"2020.
- [22] M. Jun, G. Cheng, W. Yixin, A. Xingle, G. Jiantao, Y. Ziqi et al., "Covid-19 ct lung and infection segmentation dataset 2020," 2020.
- [23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [24] E. Fernandez-Moral, R. Martins, D. Wolf, and P. Rives, "A new metric for evaluating semantic segmentation: leveraging global and contour accuracy," in *2018 IEEE intelligent vehicles symposium (iv)*. IEEE, 2018, pp. 1051-1056.
- [25] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation Coefficient metric," *PloS one*, vol. 12, no. 6, p. e0177678, 2017.
- [26] T. Trongtirakul, A. Oulefki, S. Agaian, and W. Chiracharit, "Enhancement and segmentation of breast thermograms," in *Mobile Multimedia/Image Processing, Security, and Applications 2020*, vol. 11399. International Society for Optics and Photonics, 2020, p.113990F.
- [27] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663-2674, 2018.

# Elevint: A Cloud-based Internet of Elevators

Sarah Mohammed Aljadani, Shahd Mohammed Almutairi, Saja Saeed Ghaleb, Lama Al Khuzayem

Department of Computer Science  
King Abdulaziz University  
Jeddah, Saudi Arabia

**Abstract**—With the significant growth of the number of high-rise buildings nowadays, the dependence on elevators has also increased. The issue that faces elevator passengers in case of breakdowns is the long waiting time for the arrival of the maintenance engineers to perform the repair, as the process of reporting is done manually. The safety concern increases when people are trapped. Most state-of-the-art approaches detect faults without providing means to facilitate the communication between elevator owners, maintenance companies, and engineers or notify them in case of breakdowns. Moreover, none of the proposed fault detection solutions rely on rules specified by experts in the field. This paper aims at addressing these issues by proposing a system that manages, monitors, detects faults and informs users of any faults instantly by sending notifications. Specifically, the paper proposes a mobile application, Elevint, that is, cloud-based and exploits the Internet of Things (IoT) technology. Elevint provides real-time monitoring of elevator operating conditions collected from sensors. The data is then transferred to the cloud, where faults are detected by applying rules that compare the current conditions with severity levels determined by experts. In the case that a fault is detected, elevator owners and maintenance companies are automatically notified. Moreover, through Elevint, maintenance companies can assign engineers to repair the fault and elevator owners can view and re-schedule the engineer's visit if needed. Testing our system on an elevator model shows 98% accuracy. In future, we intend to test it on real elevators to verify its applicability in practice.

**Keywords**—Internet of things; elevator; fault detection; monitoring; notification; real-time

## I. INTRODUCTION

The use of elevators has increased dramatically throughout the world, due to the congestion of population in main cities. Thus, cities have tended to increase the number of skyscrapers to accommodate the high number of populations. Successively, the reliance on elevators around the world has increased. Safety remains a critical issue in the design of elevators, since its failure may endanger people, disrupt their schedule, and threaten their lives.

Presently, when an elevator breaks down or people are trapped, there is an extensive waiting time before the relevant people are notified. This is because the process of informing the maintenance technicians is done manually. If this process is somehow automated, the maintenance team will arrive faster, and the breakdown time will be reduced significantly.

Therefore, to ensure the safety of elevator passengers, it is necessary to speed up the process of notifying the elevator owners and maintenance companies about failures. Developing notice of the breakdown in a timely manner will

not only reduce the downtime and save money, but will also allow preventive maintenance of the elevator [1].

With the advancement of technology in the 21st century, researchers in academia have considered investigating the safety control system of elevators through the Internet of Things (IoT) with the aim of moving the industry from preventative maintenance to predictive maintenance [2]. Most state-of-the-art research in this area proposes a monitoring elevator system that exploits IoT techniques to achieve the supervisory functions of early warning to reduce elevator accidents [3] while using smart sensors and cloud services [4]. Monitoring will help do analysis research to identify the main factors that cause accidents; thus, it will eventually help reduce accidents and injury rates [5].

On the other hand, elevator companies have also started to make smart elevators and are investing billions of dollars in this new technology [6]. However, it is not possible for all elevator owners to replace their old elevators with new smart ones, since there would be significant costs. This research aims at filling the gap by elevating current conventional elevators to smart elevators, through installing safety control systems that reduce downtime by analyzing the elevators' data and notifying the right people in a timely manner. Specifically, this paper proposes an IoT cloud-based elevator monitoring and fault detection mobile application which includes the following features:

- 1) Enable elevator owners and maintenance companies to monitor their elevators' operating conditions from anywhere in real-time.
- 2) Facilitate the communication between elevator owners, maintenance companies, and technicians through instant messaging services.
- 3) Detect major underlying factors for elevator faults based on some expert-defined heuristic rules.
- 4) Notify elevator owners and maintenance companies instantly about detected faults.

The remainder of this paper is organized as follows: in Section II, we review related work in this area and in Section III, we explain our proposed solution. We present our results in Section IV, followed by Section V which concludes the paper.

## II. RELATED WORK

Wang et al. [7] developed a system called EleSense, which is a framework for high-rise building structure monitoring. They used two kinds of sensing devices, the sensors are two

sets of vibrating wire strain gauge sensors attached at the inner tube and outer tube in each floor, and temperature sensors in most of the floors. A cluster head installed on each floor sends the data gathered from sensors to the base station. The base station is attached in the top of the elevator and while the elevator is moving, the base station moves with it and collects data from sensors. The results show that EleSense can significantly reduce communication costs, while providing reliable data.

Jiang et al. [8] proposed a system that provides fault detection based on real-time data from sensors that collect various parameters of running elevators, such as the weight of the lift car, the signal of portal crane, the signal of the layer precision of the elevator and the signal of safety gear. The system also enabled trapped people to communicate with maintenance staff. Moreover, the system has video and audio monitoring with Wimax (Worldwide Interoperability for Microwave Access) technology, which provides video and audio transmission based on the Internet Protocol version 4.

Zarikas and Tursynbek [9] described a system for an intelligent elevator, integrated with a smart building. The system's aim is to build a decision engine that can control the elevator's actions through AI techniques.

Salim and Akin [10] proposed an IoT-based elevator system that predicts and diagnoses errors from elevators. In this system, after troubleshooting, the data is collected from the elevators via sensors. This data is then transferred from the elevator to the control system, over the Internet to be sent to the data maintenance company to solve the problems. The system used fire sensors, low electricity and door status to detect faults. The advantage of this system is that it has a website to monitor the status of elevators, in addition to saving all elevator data and extracting reports.

Zhang et al. [11] measured the level of comfort during the elevator ride by measuring the acceleration on three directions axes x, y and z using sensors embedded in a smartphone. The elevator ride comfort level is evaluated based on the collected data, and the assessment results are uploaded to a structural health monitoring site. The results of the experiment show that their method has met the engineering requirements. Using the sensors inside the mobile device to monitor the elevator vibration may be easier and saves the time and effort of installing these sensors into a device on the elevator.

Suárez et al. [12] proposed an application that informs the maintenance company of the elevators' failures and trapped people. The application, which is developed using Microsoft Azure, implements on a cloud that is responsible for receiving data from elevators and processing it, then sending notifications accordingly. The disadvantage of this system is that it only deals with elevator faults as opposed to other systems which can also cater for management and monitoring in real-time.

Oalere et al. [13] developed a system to reduce elevators breakdowns, by early reporting of faults and diagnosis from historical data gathered from monitoring the vibration and noise using sensors. The system is divided into three layers. The first layer is the sensors layer that includes two kinds of

sensors for vibration and audio. The second layer contains the Yun microcontroller, which is an Arduino development board that analyses the signals from the sensors and helps to achieve the IoT communication. The third layer includes the web application server. They measured the results over a month, and the results showed that the elevator stops were noticeably reduced, due to notification of breakdowns. This system is very close to our proposed system but lacks some extra sensors and a monitoring camera.

Ming et al. [14] proposed an elevator safety monitoring system based on IoT technology and included audio and video monitoring. The research used sensors for vibration, acceleration, speed, diving noise, direction, floor station, elevator car door switch, voltage, temperature, and all these sensors were used to analyze data and increase safety.

Li [15] proposed a framework for an elevator security monitoring system based on cloud computing to monitor the running status of the elevator in real time. The system can provide reminders of faults, show alarm information and can classify the alarm information according to the severity.

Zhou et al. [16] introduced a remote elevator monitoring system based on IoT through real-time monitoring, fault diagnosis, alarm, and maintenance, without mentioning any details of the proposed system and results.

Huang et al. [17] proposed a warning system, which can monitor elevators failures through sensors and send the data to the remote monitoring platform. The system can send the information and remind the maintenance personnel to deal with it as soon as possible.

An et al. (a) [18] proposed an elevator monitoring system with the aim of helping managers in elevator maintenance companies to detect elevator faults early and to ensure the safety of passengers. Several sensors such as, temperature, vibration, speed, and load were exploited to capture elevators' data in real-time. This data is then sent through an industrial gateway to an InterProcess Communication (IPC) server before being sent to a SQL server through a General Packet Radio Service (GPRS) network. The web application also includes real-time video-monitoring, and a fault prediction model based on a Prognostics and Health Management (PHM) technology. The disadvantage of the system, however, is that it is a web-based application rather than a mobile application and does not provide notifications in case a fault is being detected.

An et al. (b) [19] proposed an intelligent elevator management system based on Building Information Modeling (BIM) and IoT technologies. Specifically, sensors such as temperature, humidity, speed, vibration, and load are employed to collect input data, then this data is sent through a RS485/CAN communication to the cloud for processing. The desktop application also includes real-time video monitoring as well as some functions such as, voice assistance and emergency assessment. The disadvantage of the system, however, is that it is a desktop application rather than a mobile application and does not provide means for detecting faults.

Shen et al. [20] proposed a predictive maintenance system using IoT and Machine Learning (ML) for a Permanent

Magnet Synchronous Motor (PMSM) traction elevator. First, raw data from the temperature and encoder sensors are captured then processed by Arduino. Data analysis using MATLAB is then performed on the digital data to compile it into a 30-minute data file. The data file then gets analyzed to determine which data requires predictive maintenance. The data is then classified into four categories: long, medium, short, and urgent according to predefined thresholds. Finally, this data is used for training the K-nearest Neighbor (KNN) ML model, so new data can be predicted accordingly. The accuracy of the system is 95.5%. The disadvantage of the system, however, is that it relies on two sensors' data only, and it is not a mobile application and does not provide notifications.

Guo et al. [21] described a system to monitor elevators in real-time based on multi-sensor fusion to ensure detecting common elevators faults then present the results on a website. The sensors consisted of magnetic switches, network camera, barometer, and accelerometer. The system was installed on a real elevator to test it.

Bai et al. [22] proposed a fault-prediction model based on improved PSO-BP which is optimized by an improved particle swarm optimization algorithm to enhance elevators safety using real-time data collected from the SCADA of the elevator. The convergence rate of the improved PSO-BP model is increased by 35.47%, and the prediction accuracy is improved by 49.12%

Gupta et al. [23] proposed a solution to use the elevators during COVID-19, without touching any surface to maintain sterilization and safety by deploying facial recognition software that uses pattern recognition, voice command using speech analysis method, and the body temperature to notify the people around them if someone has more than the suggested temperature. All these sensors are embedded in one device.

Table I shows a comparison between the related work according to the aim and criteria that we have developed. The first criterion (A) indicates whether the system is cloud-based or not. Secondly, (B) specifies if the system includes a fault detection model or not. Thirdly, (C) specifies if the system have video monitoring. Subsequently, (D) indicates whether the system has a mobile application or not. Finally, we determine if the system is implemented (E), available (F) or provide notifications (G). Out of the 18 methods, seven are cloud-based, fourteen systems include fault detection (some of them were predictions), nine systems include video monitoring, only three systems are mobile applications, twelve are implemented, three systems are available, and only five systems provide notifications. As can be seen from the table, all approaches include at most four features, except for Guo et al. which includes five.

As far as we know, there isn't a system that provides all features. Therefore, we aim to build an IoT cloud-based fault detection mobile application, called Elevint, where data is collected from several sensors, combining related elevators worldwide. Data will then be analyzed in the cloud to be monitored and faults can be detected using rules determined by experts. Once a fault is detected, the system will alert the

maintenance company, which in return will repair it. Thus, Elevint will include all features except for video monitoring, which we leave as future work.

TABLE I. COMPARISON BETWEEN RELATED WORK

| Research by                    | Aim                                                  | A | B | C | D | E | F | G |
|--------------------------------|------------------------------------------------------|---|---|---|---|---|---|---|
| Wang et al. (2011)             | Reduce Data Collection Delay                         | × | × | × | × | ✓ | × | × |
| Jiang et al. (2015)            | Fault Detection System                               | × | ✓ | ✓ | × | × | × | × |
| Zarikas and Tursynbek (2017)   | Control the Elevators Actions                        | × | ✓ | ✓ | × | × | × | × |
| Salim and Akin (2017)          | Fault Prediction                                     | × | ✓ | × | × | ✓ | × | × |
| Zhang et al. (2017)            | Elevator Monitoring                                  | × | ✓ | × | ✓ | ✓ | ✓ | × |
| Suárez et al. (2018)           | Inform About Failure                                 | × | ✓ | ✓ | × | ✓ | × | ✓ |
| Olalere et al. (2018)          | Remote Fault Indication                              | ✓ | ✓ | × | × | ✓ | × | ✓ |
| Ming et al. (2018)             | Elevator Monitoring                                  | × | × | ✓ | ✓ | × | × | × |
| Li (2018)                      | Elevator Monitoring                                  | ✓ | ✓ | × | × | ✓ | × | × |
| Zhou et al. (2019)             | Elevator Monitoring                                  | × | ✓ | × | × | × | × | ✓ |
| Huang et al. (2020)            | Fault Detection System                               | × | ✓ | × | ✓ | ✓ | × | ✓ |
| An et al. (a) (2021)           | Fault Prediction System                              | ✓ | ✓ | ✓ | × | ✓ | × | × |
| An et al. (b) (2021)           | Elevator Monitoring                                  | ✓ | × | ✓ | × | ✓ | × | × |
| Shen et al. (2021)             | Fault Prediction System                              | × | ✓ | × | × | ✓ | × | × |
| Guo et al. (2021)              | Elevator Monitoring & Fault Detection                | ✓ | ✓ | ✓ | × | ✓ | ✓ | × |
| Bai et al. (2021)              | Fault Prediction                                     | × | × | × | × | × | × | × |
| Gupta et al. (2022)            | Elevator Monitoring & Fault Detection                | × | ✓ | ✓ | × | × | × | ✓ |
| Our Proposed Solution: Elevint | Detection, Monitoring, Managing, Notification System | ✓ | ✓ | × | ✓ | ✓ | ✓ | ✓ |

### III. PROPOSED SOLUTION

In this section, we explain our system, Elevint, which is a cloud-based remote elevator condition monitoring system that comprises a mobile application. Elevint's system architecture is depicted in Fig. 1 Elevint aims at enabling faster repairs, preventing catastrophic breakdowns, and assisting in fault diagnosis. Specifically, Elevint provides real-time monitoring of elevator operating conditions, such as vibration, temperature, weight, light, and movement including others, through fixed sensors connected to an Arduino microcontroller. The data collected from sensors is then transferred via Wi-Fi to the cloud, where faults are detected by applying rules that compare the current conditions with severity levels determined by an expert. For example, if the temperature exceeds 60 degrees Celsius in the engine room, this indicates a possible failure in the elevator engine. Therefore, notifications are sent to elevator owners and maintenance companies to speed up the process of reporting a breakdown and to reduce the downtime of elevators.

The Android-based mobile application connects elevator owners with maintenance companies and technicians, in which messages can be exchanged between them, as well as other tasks can be performed, such as scheduling periodic maintenance and viewing maintenance history. The main feature of the mobile application is that it enables all three users, to monitor the status of the elevator in real-time i.e., the elevator's current floor, the temperature, the vibration, among others. The data gathered from the sensors is sent to the FireBase real-time cloud-hosted database, through which it gets sent to the mobile application. The mobile application is set up to show updated sensor data at 25-second intervals (changeable). Using a cloud-based architecture provides a quicker maintenance service, as the maintenance team can access the elevator data online and analyze it to find out the likely nature of the fault. For the hardware platform, our system uses a computer with Windows operating system and uses Arduino to create the circuit in which the sensors get connected. As for the software platform, the system uses Basic4Android and Visual Studio to create the application.

#### A. Hardware Components

Many hardware devices were used to produce an electric circuit connected to an Arduino Mega using the Wi-Fi to collect the real data of the elevator and send it to the application. Following is a description of some sensors that were used.

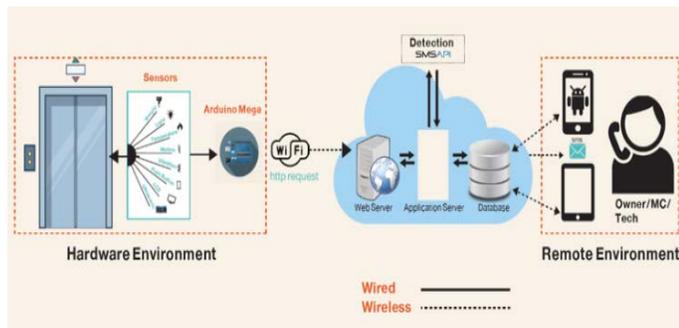


Fig. 1. Elevint System Architecture.

1) *Temperature sensor*: DS18B20 Temperature Sensor uses MAXIM's 1-wire bus protocol, which requires only 1 wire for receiving and transmitting data. The Temperature Sensor detects the temperature of the object it is attached on.

2) *Passive infraRed (PIR) sensor*: An electronic sensor that measures infrared (IR) light radiating from objects in its field of view. If the digital pulse is high (3V), it means a motion is detected, while if the digital pulse is low, it means it is idle or no motion is detected.

3) *Photo resistor light sensor*: When the value read from the photoresistor sensor module goes below a certain threshold value (determined by the developer), it means that it is dark, or the light is OFF. When the analog value from the sensor goes above the threshold value, it means the light is ON.

4) *HC-SR04 ultrasonic sensor*: The ultrasonic sensor emits an ultrasound that travels through the air. If there is an object or obstacle on its path, it will return to the module. Considering the travel time and the speed of the sound, you can calculate the distance.

5) *Vibration sensor*: Vibration is a sign of the state of the machine - no matter how accurate and unnoticed by the human senses -. Non-normal vibration of problems in the industrial machine can be detected early and repaired before machine failure occurs. Therefore, vibration analysis is used to determine the state of the equipment, location, and type of problems. Vibration sensors are sensors for measuring, displaying, and analyzing linear velocity, displacement, proximity, or acceleration.

6) *Load cell and HX711 weight sensor*: The *Load cell* senses the weight and provides an analog value to the *HX711 Load Amplifier Module*, which is an Analog to Digital Converter (ADC) that digitally converts the Load cell output.

7) *LCD display*: are used in electronic projects, as they are good for displaying information, such as the data that are collected from the sensors.

#### B. Connecting the System's Circuit

To connect the system's sensors, the circuit was made by connecting the module and sensors to the Arduino board, then the Arduino was connected to the computer using a USB cable. Fig. 2 shows the circuit connected, including the sensors and the Arduino.

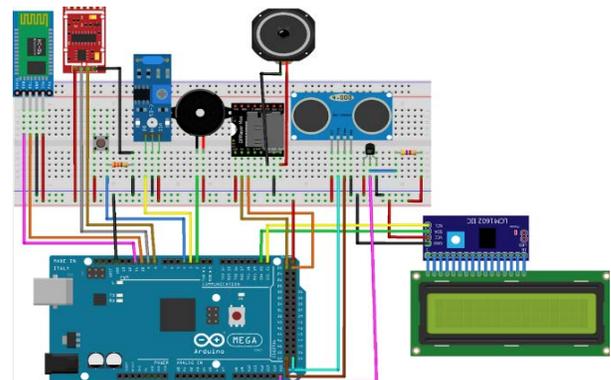


Fig. 2. The Circuit.

Table II shows the pins that are connected from the module to the pins on the Arduino of the circuit.

TABLE II. THE CIRCUIT OF ARDUINO

| Module                             |      | Arduino |
|------------------------------------|------|---------|
| <i>Temperature &amp; Vibration</i> | VCC  | + 5 V   |
|                                    | DO   | D4      |
|                                    | GND  | GND     |
| <i>Passive InfraRed</i>            | VCC  | + 5 V   |
|                                    | OUT  | PIN2    |
|                                    | GND  | GND     |
| <i>Ultrasonic</i>                  | VCC  | + 5 V   |
|                                    | Echo | 10~     |
|                                    | Trig | 11~     |
| <i>LCD Display</i>                 | VCC  | + 5 V   |
|                                    | Echo | 10~     |
|                                    | Trig | 11~     |
|                                    | GND  | GND     |

C. Elevint Application

Fig. 3 and 4 show some pages from the Elevint application. Fig. 3 (A) shows the screen where a new user is registered, and after registering, the user can view screen (B) which shows the list of functions that suits the user of the application; whether they are an elevator owner, a maintenance company, or a technician. The functions provided for an elevator owner are: adding a new elevator, monitoring elevator data in real-time, sending messages to a maintenance company and scheduling a maintenance visit. Screen (C) shows adding a new elevator, Screen (D) provides a view of the elevator’s real data, Screens (E-F) shows how to schedule a maintenance visit. In Fig. 4, screens (G-K) show how the user can view the maintenance history in detail. Screen (L) shows the message function which allows easy communication between the user and the maintenance company.

D. Detection Rules

The rules for detecting the various faults are of the form:

if condition then action(s) {else action(s)}

Note that the part between curly brackets indicates that it can be excluded. If more than one action is included, then actions will be numbered. Tables III to VI shows our rules for detecting a number of faults. The first row in each table indicates the goal of the detection rule, the second row specifies the used sensors, and the third row shows the detection rule. Note that these detection rules are not fixed rules, but rather heuristic gathered from experts in the field.

The threshold values used in the rules were determined through the performed tests on the elevator model.

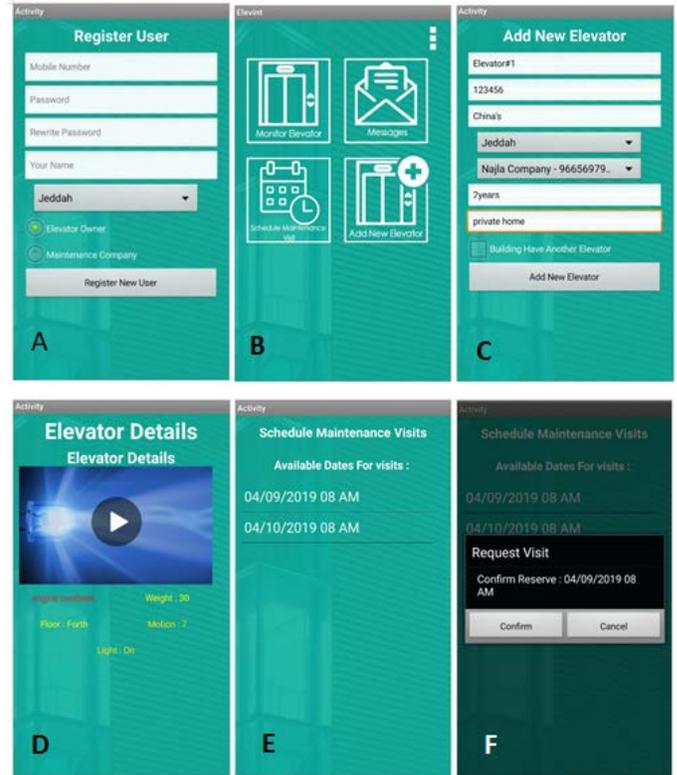


Fig. 3. App Pages for Maintenance Company Users - Part 1.

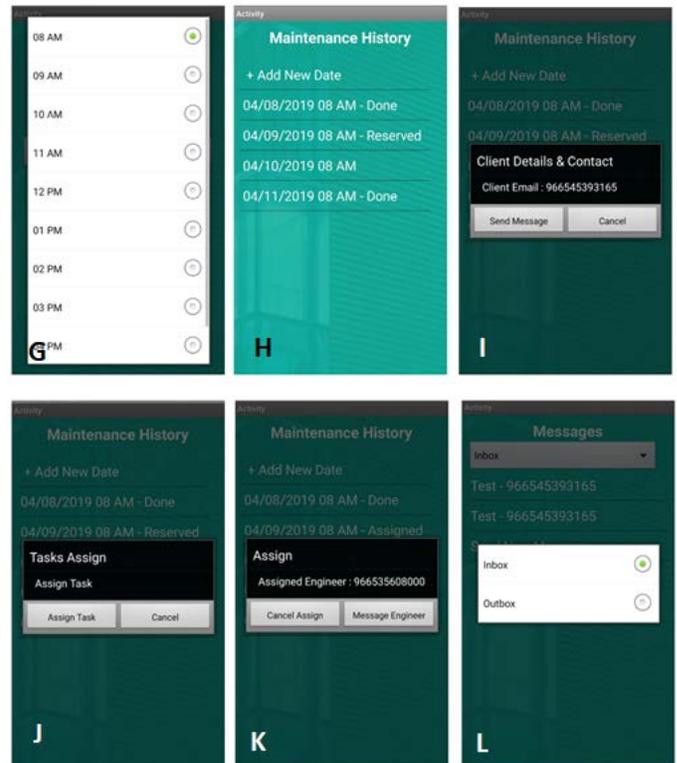


Fig. 4. App Pages for Maintenance Company Users - Part 2.

Table III demonstrates rule #1, which detects if the elevator's engine is overheated. For instance, when the temperature exceeds 60 degrees Celsius in the engine room, it indicates the possibility of a fault in the elevator engine and therefore, the LCD screen and the Monitor Elevator screen on the Elevint application will show the exact temperature, and an SMS will be sent to the owner of the elevator indicating that an overheat engine fault is possibly detected in an elevator, giving its serial number. Note that notification\_rule#6 is demonstrated in Table VII and will be explained in the next section. Otherwise, if the temperature of the engine is a normal value, it will appear on the LCD screen and will be displayed in the Elevint application. We have tested "overheated engine" faults by directing a lighter on the temperature sensor and the result was an SMS message sent to the mobile number of the owner registered on Elevint.

TABLE III. DETECTION RULE #1

| <i>detection_rule#1: Overheat Engine</i> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Goal</b>                              | Detect a possible fault in the elevator's engine (e.g., fire)                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
| <b>Sensors</b>                           | Temperature                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <b>Rule</b>                              | <p><b>if</b> <i>temperature</i> &gt; 60 °<br/> <b>then</b><br/>                     1) <i>Print the temperature on the LCD screen</i><br/>                     2) <i>Display the temperature in the application</i><br/>                     3) <i>Send SMS notification to users according to notification_rule#6 with fault = 'engine overheat'</i><br/> <b>else</b><br/>                     1) <i>Print the temperature on the LCD screen</i><br/>                     2) <i>Display the temperature in the application</i></p> |

Table IV shows rule #2, which detects if there are people trapped in an elevator. For instance, if Motion = High and Light = Off, that means there are people trapped in the elevator. Thus, the LCD screen and in the Elevint application will show: "Last Motion: 3 seconds ago", and Lighting: On. An SMS will also be sent to the elevator owner indicating that people may be trapped in the elevator with its serial number. Otherwise, the data will be displayed on the LCD screen and in the Elevint application. We have tested faults from type "Trapped People" by turning off the light and doing movements in front of the motion sensor and the result was an SMS message sent to the mobile number of the owner registered on Elevint.

Table V demonstrates rule #3, which detects if the elevator is stuck between floors. For instance, if the distance between floors is detected to be 15.5 cm and the vibration is 0, then the LCD screen and the Monitor Elevator screen on the Elevint application will show the message: "Floor: elevator is stuck", and an SMS will be sent to the owner of the elevator indicating that the elevator, giving its serial number, may be stuck. Otherwise, the data will be displayed on the LCD screen and on the Elevator Monitor in the Elevint application. We have tested "elevator stuck" faults by moving the elevator and stopping it between two floors and the result was an SMS message being sent to the mobile number of the owner registered on Elevint.

TABLE IV. DETECTION RULE #2

| <i>detection_rule#2: Trapped People</i> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Goal</b>                             | Detect possible trapped people in the elevator                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <b>Sensors</b>                          | Motion and Light                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <b>Rule</b>                             | <p><b>if</b> <i>motion = high &amp;&amp; light = OFF</i><br/> <b>then</b><br/>                     1) <i>Print last detected motion timing on the LCD screen</i><br/>                     2) <i>Display last detected motion timing in the application</i><br/>                     3) <i>Print light status on the LCD screen</i><br/>                     4) <i>Display light status in the application</i><br/>                     5) <i>Send SMS notification to users according to notification_rule#6 with fault = 'trapped people in the elevator'</i><br/> <b>else</b><br/>                     1) <i>Print last detected motion timing on the LCD screen</i><br/>                     2) <i>Display last detected motion timing in the application</i><br/>                     3) <i>Print light status on the LCD screen</i><br/>                     4) <i>Display light status in the application</i></p> |

TABLE V. DETECTION RULE #3

| <i>detection_rule#3: Stuck Elevator</i> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Goal</b>                             | Detect if the elevator is stuck between floors                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>Sensors</b>                          | Distance and Vibration                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| <b>Rule</b>                             | <p><b>if</b> <math>(!(distance \geq 0 \text{ cm} \ \&amp;\&amp; \ distance \leq 15 \text{ cm}) \parallel (distance &gt; 15 \text{ cm} \ \&amp;\&amp; \ distance \leq 25 \text{ cm}) \parallel (distance &gt; 25 \text{ cm} \ \&amp;\&amp; \ distance \leq 35 \text{ cm}) \parallel (distance &gt; 35 \text{ cm} \ \&amp;\&amp; \ distance \leq 45 \text{ cm} \ \&amp;\&amp; \ vibration = 0)</math><br/> <b>then</b><br/>                     1) <i>Print 'elevator is stuck' on the LCD screen</i><br/>                     2) <i>Display 'elevator is stuck' in the application</i><br/>                     3) <i>Send SMS notification to users according to notification_rule#6 with fault = 'stuck elevator'</i><br/> <b>else</b><br/>                     1) <i>Print current elevator floor on the LCD screen</i><br/>                     2) <i>Display current elevator floor in the application</i></p> |

Table VI demonstrates rule #4, which detects if the elevator is overloaded. For instance, if the weight is 1200 g, then the LCD screen that is attached on the elevator and the Monitor Elevator screen on the Elevint application will show the message: "Weight: elevator overloaded", and an SMS will be sent to the owner of the elevator indicating that the elevator, giving its serial number, is overloaded. Otherwise, the data will be displayed on the LCD and on the Elevator Monitor in the Elevint application. We have tested faults from type "overloaded elevator" by putting a weight on the weight sensor and the result was an SMS message sent to the mobile number of the owner registered on Elevint.

E. System Notification Rules

In our application, we send SMS notifications to users when certain events take place according to the rules of the form:

when event then action(s)

TABLE VI. DETECTION RULE #4

| <b>detection_rule#4: Overloaded Elevator</b> |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |
|----------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Goal</b>                                  | Detect if the elevator is overloaded                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Sensors</b>                               | Weight                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <b>Rule</b>                                  | <p><b>if</b> <i>weight</i> &gt;= 1000 g<br/> <b>then</b><br/>                     1) <i>Print lift overloaded on the LCD screen</i><br/>                     2) <i>Display current weight in the application</i><br/>                     3) <i>Send SMS notification to users according to notification_rule#6 with fault = 'elevator is overloaded'</i><br/> <b>else</b><br/>                     1) <i>Print current weight on the LCD screen</i><br/>                     2) <i>Display current weight in the application</i></p> |

where an event is an event generated by the various users of the application, such as selecting a choice, sending or receiving messages, booking an appointment, or assigning tasks, and action is a call to the function *sendSMS*, which takes two arguments, the receiver of the SMS and the message itself. Table VII below shows one notification rule, rule #6 which captures the case when a fault is detected.

TABLE VII. NOTIFICATION RULE #6

| <b>notification_rule#6: Detect Fault</b> |                                                                                                                                                                                                                                                                                                                                                                                |
|------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Goal</b>                              | Informing a maintenance company and an elevator owner about a detected fault.                                                                                                                                                                                                                                                                                                  |
| <b>Event</b>                             | <i>a fault is detected</i>                                                                                                                                                                                                                                                                                                                                                     |
| <b>Action</b>                            | 1) <i>sendSMS</i> (elevator owner, "Hi <i>owner_name</i> , a fault has been detected in the elevator with the serial number: <i>serial_no</i> , possible fault: <i>fault</i> ."<br>2) <i>sendSMS</i> (maintenance company, "Hi <i>mainComp_name</i> , a fault has been detected in the elevator with the serial number: <i>serial_no</i> , possible fault: <i>fault</i> ."<br> |

#### IV. EXPERIMENTAL RESULTS

We tested our fault detection rules by creating manually 10 faults for each of the 5 rules and calculating the results using the classification metrics from the information retrieval using the formulas below:

$$\text{precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{recall} = \frac{TP}{TP+FN} \quad (2)$$

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (3)$$

$$\text{accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

where *TP* is the number of correctly detected real faults, *FP* is the number of detected faults which were wrong (i.e., not fault), *FN* is the number of undetected faults which were true (i.e., failed to detect) and *TN* is the number of undetected faults which were not fault (i.e., normal case).

Fig. 5 shows the results for all five detection rules, in which, the precision, calculated from (1), was 1 for all rules, while the recall, calculated from (2), was 1 in two rules only. For rules 1, 4 and 5 the recall scores were 0.86, 0.8, 0.9 respectively. The average precision for all rules is 1 and the average recall is 0.91. The F1 measure from (3) was also calculated and the result is 0.95. Finally, the accuracy from (4) was calculated and the result is 0.98.

Fig. 6 shows the testing of Detection Rule #4 and Fig. 7 shows the elevator model that was used to evaluate Elevint.

**Detection Rules Results**

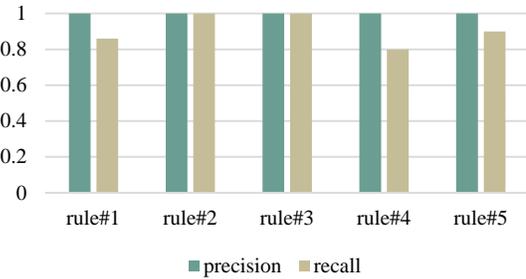


Fig. 5. Elevint's Detection Rules Results.

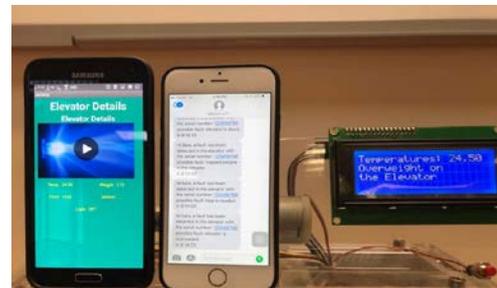


Fig. 6. Testing Detection Rule #4.



Fig. 7. Elevint's Model.

Fig. 8 shows the SMS that was sent to the owner due to the testing of Detection Rule #3.

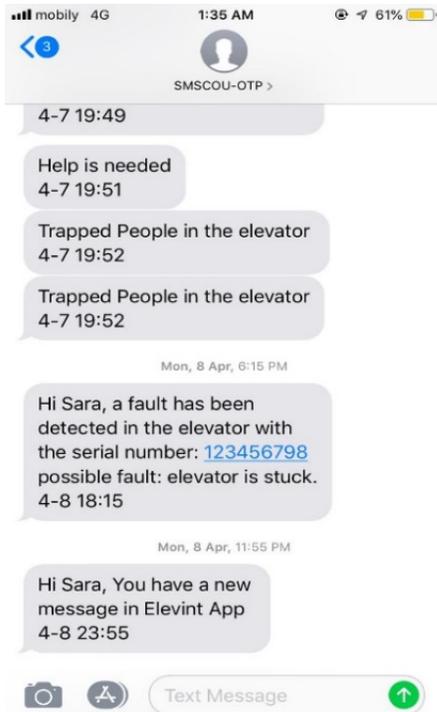


Fig. 8. Testing Detection Rule #3.

## V. CONCLUSION

Undoubtedly, elevator failures can cause significant delays and threats to humans' lives if not handled quickly and efficiently. A main cause of the delay is the manual approach taken to report elevator breakdowns. The problem with most state-of-the-art approaches is that they concentrate on monitoring and detecting faults without providing automatic means for notifying users when these faults happen. Moreover, their fault detection mechanisms do not rely on rules specified by experts in the field. Furthermore, none of the previous approaches offered means for managing elevators and facilitating the communication between elevator owners, maintenance companies and engineers. As opposed to previous approaches, this paper has proposed Elevint, an IoT, cloud-based mobile application that provides managing, monitoring, fault detection and notification services. Specifically, through Elevint, elevator owners can request a maintenance visit and monitor their elevators' data in real-time, while maintenance companies can assign engineers to specific faults or scheduled visits, and view elevators' data and maintenance history. Moreover, all users can communicate with each other through Elevint via messaging. Furthermore, faults are detected by applying heuristic rules that compare the current conditions with severity levels determined by experts. Finally, when a fault is detected, elevator owners and maintenance companies are automatically notified. Testing Elevint on an elevator model showed promising results. In the future, we aim to increase the number of detected faults by adding more detection rules, and to test our system on real elevators. We also intend to improve the

system further by offering fault prediction rather than fault detection using AI techniques.

## REFERENCES

- [1] Perkel, C. (2016). Trapped in a Stuck Elevator? The Problem is Worsening in Canada, Experts Warn. Retrieved from TheStar: <https://www.thestar.com/news/canada/2016/07/21/trapped-in-a-stuck-elevator-the-problem-is-worsening-in-canada-experts-warn.html> (last visited: October 22, 2021).
- [2] Hou, K., Tao, S., But, M., and Cho, K. (2016). Smart Elevator. 1-18.
- [3] Lueth, K. (2016). The 10 Most Popular Internet of Things Applications Right Now. Retrieved from IoT Analytics: <https://iot-analytics.com/10-internet-of-things-applications/> (last visited: October 22, 2021).
- [4] Solanki, K. (2018). A Comprehensive Study on Smart City: Concept and Limiting Factors. International Journal of Advanced Research in Computer Science, 9(2).
- [5] Shafique, M., & Rafiq, M. (2019). An Overview of Construction Occupational Accidents in Hong Kong: A Recent Trend & Future Perspective. Applied Sciences, 9(10), 2069.
- [6] Propmodo. (2015). IoT Connected Elevators May Save Office Workers Years of Waiting, Retrieved From: <https://www.propmodo.com/iot-connected-elevators-may-save-office-workers-years-of-waiting/> (last visited: November 30, 2021).
- [7] Wang, F., Wang, D., & Liu, J. (2011). Utilizing Elevator for Wireless Sensor Data Collection in High-Rise Structure Monitoring. Paper Presented at the Proceedings of the Nineteenth International Workshop on Quality of Service, 1-9.
- [8] Jiang, H., Shi, Y., & Qi, L. (2015). Design of Elevator Monitoring and Alarm System Based on WiMAX. Proceedings of the 2015 International Conference on Electrical, Computer Engineering & Electronics, 92-96.
- [9] Zarikas, V., & Tursynbek, N. (2017). Intelligent Elevators in a Smart Building. Paper presented at the Future Technologies Conference 2017.
- [10] Salim, O., & Akin, E. (2017). IoT Application for Fault Diagnosis and Prediction in Elevators. International Journal of Innovative Research in Science, Engineering & Technology, 6(4), 5737-5742.
- [11] Zhang, Y., Sun, X., Zhao, X., & Su, W. (2018). Elevator Ride Comfort Monitoring and Evaluation Using Smartphones. Mechanical Systems & Signal Processing, 105, 377-390.
- [12] Suárez, A., Parra, O., & Forero, J. (2018). Design of an Elevator Monitoring Application Using Internet of Things. International Journal of Applied Engineering Research, 13(6), 4195-4202.
- [13] Olalere, I., Dewa, M., & Nleya, B. (2018). Remote Condition Monitoring of Elevator's Vibration and Acoustics Parameters for Optimized Maintenance Using IoT Technology. Paper presented at the 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE).
- [14] Ming, Z., Han, S., Zhang, Z., & Xia, S. (2018). Elevator Safety Monitoring System Based on Internet of Things. International Journal of Online Engineering, 14(8), 121-
- [15] Li, J. (2018). Design and Implementation of Elevator Internet of Things Security Control System based on Cloud Computing.
- [16] Zhou, Y., Wang, K., & Liu, H. (2018). An Elevator Monitoring System Based on The Internet of Things. Procedia Computer Science, 131, 541-544.
- [17] Huang, Q., Cao, J., & Sun, R. (2020, August). Design and Implementation of an Elevator Power Failure Warning System. In *Journal of Physics: Conference Series* (Vol. 1621, No. 1, p. 012050). IOP Publishing.
- [18] An, Z., Liu, S., Bai, D., Wang, N., & Yu, X. (2021). Intelligent Monitoring System of Elevator Internet of Things. In *IOP Conference Series: Earth & Environmental Science* (Vol. 791, No. 1, p. 012127). IOP Publishing.
- [19] An, Z., Bai, D., Huang, Y., Ning, W., Deng, Y., Gan, N., & Liu, S. (2021). Building Elevator Safety Monitoring System Based on the BIM Technology. In *Journal of Physics: Conference Series* (Vol. 1939, No. 1, p. 012026). IOP Publishing.

- [20] Shen, L. J., Lukose, J., & Young, L. C. (2021). Predictive Maintenance on an Elevator System using Machine Learning. *Journal of Applied Technology & Innovation* (e-ISSN: 2600-7304), 5(1), 75.
- [21] Guo, Y., Liu, Y., Zhang, X., & Wang, G. (2021). The Real-Time Elevator Monitoring System Based on Multi-sensor Fusion. In *Journal of Physics: Conference Series* (Vol. 2010, No. 1, p. 012182). IOP Publishing.
- [22] Bai, D., An, Z., Wang, N., Liu, S., & Yu, X. (2021). The Prediction of the Elevator Fault Based on Improved PSO-BP Algorithm. In *Journal of Physics: Conference Series* (Vol. 1906, No. 1, p. 012017). IOP Publishing.
- [23] Gupta, S., Tyagi, S., & Kishor, K. (2022). Study and Development of Self Sanitizing Smart Elevator. In *Proceedings of Data Analytics & Management* (pp. 165-179). Springer, Singapore.

# Moving Object Detection over Wireless Visual Sensor Networks using Spectral Dual Mode Background Subtraction

Ahmed M. AbdelTawab<sup>1\*</sup>, M.B. Abdelhalim<sup>2</sup>, S.E.D. Habib<sup>3</sup>

Electronics and Communications Department, Faculty of Engineering, Misr University for Science & Technology MUST, Giza, Egypt<sup>1</sup>

College of Computing and Information Technology, Arab Academy of Science and Technology and Maritime Transport, Cairo, Egypt<sup>2</sup>

Electronics and Communications Department, Faculty of Engineering, Cairo University, Giza, Egypt<sup>3</sup>

**Abstract**—Wireless Visual Sensor Networks (WVSN) play an essential role in tracking moving objects. WVSN's key drawbacks are storage, power, and bandwidth. Background subtraction is used in the early stages of target tracking to extract moving targets from video images. Many standard methods of subtracting backgrounds are no longer suitable for embedded devices because they use complex statistical models to manage small changes in lighting. This paper introduces a system based on the Partial Discrete Cosine Transform (PDCT), reducing the vast dimensions of processed data while retaining most of the important information, thereby reducing processing and transmission energy. It also uses a dual-mode single Gaussian model (SGM) for accurate detection of moving objects. The proposed system's performance is to be assessed using the standard CDnet 2014 benchmark dataset in terms of detection accuracy and time complexity. Furthermore, the suggested method is compared to previous WVSN background subtraction methods. Simulation results show that the proposed method consistently has 15% better accuracy and is up to 3 times faster than the state-of-the-art object detection methods for WVSN. Finally, we showed the practicality of the suggested method by simulating it in a sensor network environment using the Contiki OS Cooja Simulator and implementing it in a real testbed using Cortex M3 open nodes of IOT-LAB.

**Keywords**—Background subtraction; discrete cosine transform; embedded camera networks; Gaussian mixture models; wireless visual sensor networks

## I. INTRODUCTION

Wireless sensor networks (WSNs), which are made up of thousands of scalar sensors nodes that are spatially distributed and wirelessly communicated, have attracted researchers' interest [1]. Small and low-power CMOS cameras and microphones are used in Wireless Visual Sensor Networks (WVSNs), which can collect visual cues from the environment. The WSN's capabilities are being expanded to include sophisticated environmental monitoring, advanced health care delivery, traffic avoidance, fire prevention, and monitoring, as well as object tracking, and modern surveillance systems [2]. WVSN has focused on military, commercial traffic management, and precision agriculture surveillance applications [3]. Three major problems make WVSNs lack vision processing capability. First, sensor nodes' visual

processing capability, second, memory storage constraints for sensor nodes and Finlay; communication of large volumes of image data. However, maximising network lifespan while processing huge volumes of multimedia data while following application-specific QoS requirements such as latency, packet loss, bandwidth, and throughput is a challenge. In addition to developing energy-sensitive multimedia processing algorithms and infrastructures, it is also necessary to establish efficient communication strategies [3].

Object detection is the first and most critical step in target tracking [4]. Robust object detection is typically the dominant consumer of processing and resources, where the moving targets are extracted from the video frames to perform further high-level processing. Lighting changes, shifting backgrounds, artificial or fast motion, and occlusion make accurate foreground object segmentation challenging [5]. The major methodologies for completing the object detection task include optical flow [6], frame differencing [7], and background subtraction [8].

Background subtraction is a standard and consistent method for detecting moving foreground that involves subtracting the background model from the current frame and changing the background model on a regular basis to remove the effects of illumination and inappropriate events. This method is extensively used for motion detection tasks in dynamic scenarios. In practice, basic techniques like mixture of Gaussians (MOG) [9], KDE [10], codebook [11], and ViBe [12] are employed for real applications. Despite the accuracy and efficiency of the MoG [9], the evaluation in [13] demonstrates that MoG can only handle three frames per second on the Blackfin DSP camera nodes with a low image resolution frame size of  $320 \times 240$ . The need to update the MoG probability distribution parameters accounts for the long computation time of MoG.

This work aims to investigate the development of moving object detection over WVSN. The Discrete Cosine Transform (DCT) [14] is a frequently utilised image compression technique over WVSN [15, 16]. The DCT algorithm converts signals from the spatial domain to a frequency domain representation. We apply the DCT to minimise the dimensionality of the background subtraction problem while

\*Corresponding Author.

maintaining accuracy. The following are the contributions made by this paper:

- A new compression-based background subtraction called Spectral Dual Mode Background Subtraction (SDMBS) uses Partial Discrete Cosine Transform (PDCT) [15] (for dimensionality reduction) and Dual mode SGM [17] (for accuracy) to model the background and distinguish the foreground from the background.
- We implement our approach and compare it to MoG and other compressed-based MoG methods to demonstrate the computational efficiency of our suggested methods. According to the results, our method is up to 10 times faster than the original MoG and three times faster than the compressed-based MoG.
- To demonstrate the algorithm's ability to work in wireless sensor network environments, we simulated and realised the proposed SDMBS in a Cooja network simulator and on the IOT-LAB M3 board.

The rest of the paper is organized as follows. We first present the related work in Section II. We then present a detailed account of the proposed SDMBS approach in Section III. Section IV discusses the simulation results and performance evaluation in detail. Section V draws the paper's conclusion.

## II. RELATED WORK

### A. Object Detection in WVSN

In visual sensor networks, the cost of data communication is usually far higher than the cost of image processing. As a result, traditional object detection methodologies are ineffective for monitoring and surveillance applications; instead, the image raw data is sent to the sink node, where detection methods are used to determine the moving object. Alternative approaches are to either compress the image at the sensor node and apply object detection at the sink node after decompression, or process the frame before transmission and transmit the useful information or features for further analysis at the sink node. Compression can be applied using Compressed Sensing (CS), wavelet, or DCT. In the second approach; frame processing is applied either on raw image data or compressed domain to further reduce processing complexity at the sensor node. The compressed data is already computed and has less storage space than the raw image frame. The two approaches are briefly reviewed in this section.

1) *Compressed data*: According to Robust Primary Component Analysis (RPCA) [18], DECOLOR [19], the basic concept of low-order factorization structures and sparse factorization is to divide a given matrix of acquired frames into background and sparse foreground by outliers. The goal of Compressed Sensing CS (low-rank BS) [20] is to send a compressed image to the base station using Compressed Sensing (CS) [13] and then use Orthogonal Matching Pursuit (OMP) [21] to rebuild the image at the receiver end. The authors of [22] proposed a CS-based detection approach that

uses CS measurements of a moving object to reconstruct the foreground in a video.

2) *Processed data*: Because the video to be sent in surveillance applications is generally static, a resource-constrained environment like WVSN does not require the transmission of the entire video. The video can be processed using a compression-based background removal technique to recognise moving objects and send only the foreground data to the monitoring location to save energy and bandwidth. A method for sending image portions instead of the whole image as described in [23]. It ensures that the sink node receives the bare minimum of image content, as assessed by in-node energy consumption and reconstructed picture peak signal-to-noise ratio (PSNR). The image processing block (Running Gaussian Average technique for object extraction and DWT for ROI transmission) operates at a high frequency to facilitate rapid processing and is only engaged by a separate network processor when images need to be analysed [24]. Because it runs continually, the network processor block is designed to operate at a low frequency. The suggested approach for image processing and communication requires relatively little energy, as evidenced by practical test and simulation results. To save transmission energy, Nandhini et al. [25] propose a method for detecting objects with fewer measures that combines a mean measurement differencing approach with an adaptive threshold strategy.

CS-based background subtraction is measured based on the node before object information is sent, reducing complexity in terms of power, storage, and bandwidth. CSMOG [26] applies MOG [9] to low-rate CS measurements. CSMOG [26] is based on the idea of reducing the number of dimensions in data while still capturing the majority of the information via a random projection matrix. The CSMOG method is consistently superior, up to 6x faster, and uses significantly fewer resources than the standard method, according to real-time requirements. The DWT-based CS object identification framework [27] uses a simple measurement matrix termed the deadweight tonnage block diagonal matrix to refine the pixel-based foreground following the block-based foreground recognition phase in the first stage. The averaging approach using the Adaptive Threshold Technology (MMDATS) in [25] is based on the framework for robust subspace learning. The OMP approach is used to reconstruct the object from foreground measurements. Due to its excellent directional selectivity and shift-invariance, [28] uses a motion segmentation algorithm based on interframe differentiation using the complex Daubechies wavelet transform in the wavelet compression domain.

To reduce the storage space and time required, a background statistical subtraction approach [29] based on motion segmentation in the compression transform domain using Wavelet has been proposed. A good observation was made in 8x8 blocks using the DCT coefficients of the pre-coded JPEG image [22]. They developed a background subtraction strategy that properly depicts the background model over time using competing Hidden Markov Models (HMM). Three techniques for modelling the background directly from the compressed video are presented in [30].

Moving average, median, Gauss blending. These methods use the DCT coefficient (including the AC coefficient) to characterise the background at the block level, and then update the DCT coefficient to match the background. Popa et al. [31] use low DCT compressed area processing to simulate the background. Processing at the block level instead of the pixel level reduces the number of simulation parameters by almost one-third. They also reduce the number of coefficients per block from 64 to 16 while retaining segmentation quality. In the DCT domain, Ye et al. [32] evaluated the background stability and separability of objects. The suggested method restores the target by suppressing the background coefficients by modelling the background as a single Gaussian model for each frequency point. A quaternion-DCT for infrared target recognition is presented by [33]. This approach shows how to create a quaternion with two-directional features (motion feature and kurtosis feature). The QDCT drawing feature acts as a unique signature that helps solve problems when finding small targets. To reduce complexity and simplify hardware implementation, Manimozhi et al. [34] employed a diagonal matrix of binary substitution blocks as the measurement matrix for both DCT-based and DWT-based CS procedures.

According to related research, a large volume of video is required, as well as a significant amount of storage space and processing time for the segmentation method. Compression-based processing is recommended for restricted WVSNs to address the above issues. As a result, we'll describe a motion segmentation method using the DCT in the compression domain based on statistical background subtraction. The dual-mode SGM-based background subtraction technique recognises just the foreground blocks of the discrete cosine transform's detailed component to reduce processing complexity. Then, adjust the foreground block to recognise the foreground object. The foreground block is moved to the sink side for rebuilding and tracking. In Fig. 1, the proposed SDMBs (page size = 4) is compared to the original MOG [9] and with block measurements based on compressed sensing (CSMOG) [26].

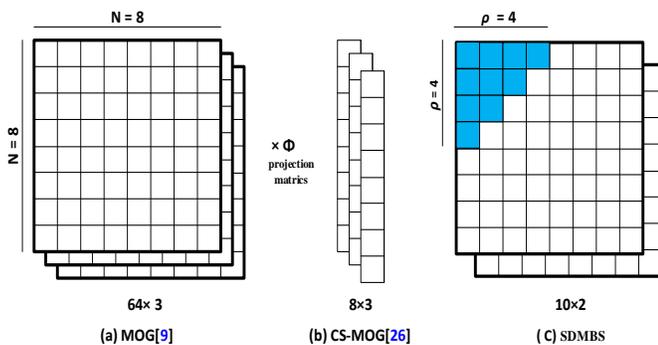


Fig. 1. Block Computation for (a) MOG, (b) CS-MOG and (c) SDMBs.

### III. PROPOSED SYSTEM MODEL

We first describe the steps of Spectral Dual-Mode Background Subtraction (SDMBs), then justify the use of dual-mode SGM (D-SGM) on top of the reduced dimension data PDCT. Fig. 2 depicts the proposed SDMBs algorithm's block diagram as well as the network topology.

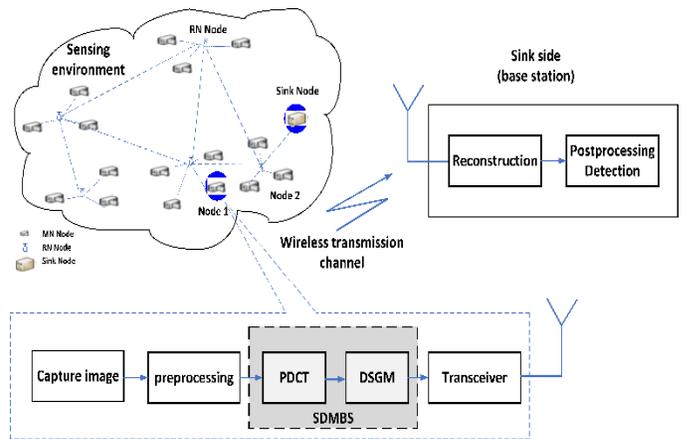


Fig. 2. The Proposed Block Diagram for WWSN-based Object Detection.

#### A. Network Model

We are considering randomly deploying WWSN nodes in the surveillance field. Each WWSN node is constrained in terms of process and memory resources. The WWSN system model is composed of  $N$  visual sensor nodes, Relaying Nodes (RNs), one or more Monitoring Node or Sensor node (MNs), and a Sink Node (SN) [23]. Each sensor node  $i$  is thought to be in a 'wakeup' state according to a unique duty cycle  $\beta_s \in [0, 1]$  during a period  $t_s$  to successfully send an image via the network. Thus, it avoids any conflicts induced by two or more nodes simultaneously broadcasting image data. Thus, each sensor is awake for a length of time  $\beta_s t_s$  and sleeps for a length of time  $(1 - \beta_s) t_s$ . The frame count is set to zero when a sensor node enters a 'wake up' condition.

#### B. Pre-Processing

Simple spatial Gaussian filtering and median filtering are used to suppress salt and pepper and Gaussian noise in images captured during the preprocessing step [27]. The filtered frame is then divided into equal-sized blocks, with the SDMBs algorithm applied to each block separately. This can be done in parallel, further reducing computation time.

#### C. Discrete Cosine Transform (PDCT)

As seen in Fig. 2, each video frame is subsequently divided into  $8 \times 8$  blocks. After that, each block is subjected to DCT. Each  $8 \times 8$  DCT block is represented by the first ten low-frequency DCT components. The partial DCT has the advantage of compressing an  $8 \times 8$  block into 10 samples, which is useful for WSNs with limited resources. Although the rest of the data is sparse, the DCT DC-coefficient stays concentrated in the series' upper left corner. Compressed sensing CS [25] requires a sparse value.

#### D. Dual Mode Signal Gaussian Model (DM-SGM)

To deal with the inaccuracies that come from modelling the scene using SGMs [35], a dual-mode SGM with age [17] is utilised. While still learning the background reliably, this model safeguards the background model from foreground and noise contamination. The compressed domain PDCT low frequency components are subjected to DM-SGM to identify whether or not the image block contains a moving target. Here, the Gaussian parameter for each grid is computed. Mean,

variance, and age are then used to model the background, determining and updating the foreground blocks. There are two models for each block; appearance background models and candidate background models. The candidate background model is ineffective until its age exceeds that of the apparent background model. This dual-mode SGM differs from two-version Gaussian combination models (GMM) [9] in that, with a bi-modal GMM, the foreground facts could still contaminate the history. However, with the dual-mode SGM approach, this is no longer the case.

The two models are switched at this point. At the end, the foreground blocks are determined and applied to the pixel refining stage to detect the pixels containing the target within the foreground block, according to the flowchart in Fig. 3.

The group of pixels in grid  $i$  at time  $t$  is denoted as  $\mathcal{G}_i^{(t)}$ , the number of pixels in  $\mathcal{G}_i^{(t)}$  as  $|\mathcal{G}_i^{(t)}|$ , and the observed pixel intensity of a pixel  $j$  at time  $t$  as  $I_j^{(t)}$ , and the mean  $\mu_i^{(t)}$ , the variance  $\sigma_i^{(t)}$ , and the age  $\alpha_i^{(t)}$  of the SGM model applied to  $\mathcal{G}_i^{(t)}$  is updated as

$$\mu_i^{(t)} = \frac{\tilde{\alpha}_i^{(t-1)}}{\tilde{\alpha}_i^{(t-1)}+1} \tilde{\mu}_i^{(t-1)} + \frac{1}{\tilde{\alpha}_i^{(t-1)}+1} M_i^{(t)} \quad (1)$$

$$\sigma_i^{(t)} = \frac{\tilde{\alpha}_i^{(t-1)}}{\tilde{\alpha}_i^{(t-1)}+1} \tilde{\sigma}_i^{(t-1)} + \frac{1}{\tilde{\alpha}_i^{(t-1)}+1} V_i^{(t)} \quad (2)$$

$$\alpha_i^{(t)} = \tilde{\alpha}_i^{(t-1)} + 1 \quad (3)$$

$$M_i^{(t)} = \frac{1}{|\mathcal{G}_i|} \sum_{j \in \mathcal{G}_i} I_j^{(t)} \quad (4)$$

$$V_i^{(t)} = \max_{j \in \mathcal{G}_i} (\mu_i^{(t)} - I_j^{(t)})^2 \quad (5)$$

$$V_i^{(t)} = (\mu_i^{(t)} - M_i^{(t)})^2 \quad (6)$$

DM-SGM [17] uses another SGM as a prospective background model. At this point, the candidate background model is rendered ineffectual until it reaches the same age as the apparent background model, at which point the two models are exchanged. We update the mean, variance, and age of the candidate background model and the apparent background model at time  $t$  for grid  $i$ ,  $\mu_{C,i}^{(t)}$ ,  $\sigma_{C,i}^{(t)}$ , and  $\alpha_{C,i}^{(t)}$  and  $\mu_{A,i}^{(t)}$ ,  $\sigma_{A,i}^{(t)}$ , and  $\alpha_{A,i}^{(t)}$ , respectively, according to (1), (2), and (3), if

$$(M_i^{(t)} - \mu_{A,i}^{(t)})^2 < \theta_s \sigma_{A,i}^{(t)} \quad (7)$$

Where  $M_i^{(t)}$  is the observed mean and  $\theta_s$  is a threshold parameter. Also, we update  $\mu_{C,i}^{(t)}$ ,  $\sigma_{C,i}^{(t)}$ , and  $\alpha_{C,i}^{(t)}$ , according to (1), (2), and (3).

If condition (7) is violated, and if the observed mean matches the candidate background model, then

$$(M_i^{(t)} - \mu_{C,i}^{(t)})^2 < \theta_s \sigma_{C,i}^{(t)} \quad (8)$$

If none of the conditions hold, we start the candidate background model with the current observation. Only one of the two models is altered when this process is used, while the other is left alone. If the candidate's age exceeds the apparent meaning, the two backdrop models for the grid are swapped after updating.

$$\alpha_{C,i}^{(t)} > \alpha_{A,i}^{(t)} \quad (9)$$

Once the candidate is exchanged, the background model is initialised. Finally, an apparent background model is solely employed to determine foreground pixels, as stated in Section E. preventing the background model from being distorted by the foreground data that represents the object.

The candidate background model, rather than the apparent background model, learns the foreground data in the dual-mode SGM, preventing the background model from being distorted by the foreground data that represents the moving object in the frame. So, the models are changed and the correct background model is chosen if the candidate background model's age is greater than the apparent background models.

### E. Pixels Refining

A foreground block contains both foreground and background pixels. Each video frame contains a large number of background blocks. As a result, we just need to focus on the small number of foreground blocks. To detect which pixels in a foreground block are indeed foreground, a basic background learning technique for each block is created. If we classify a pixel  $j$  in a group  $i$  as a foreground pixel,

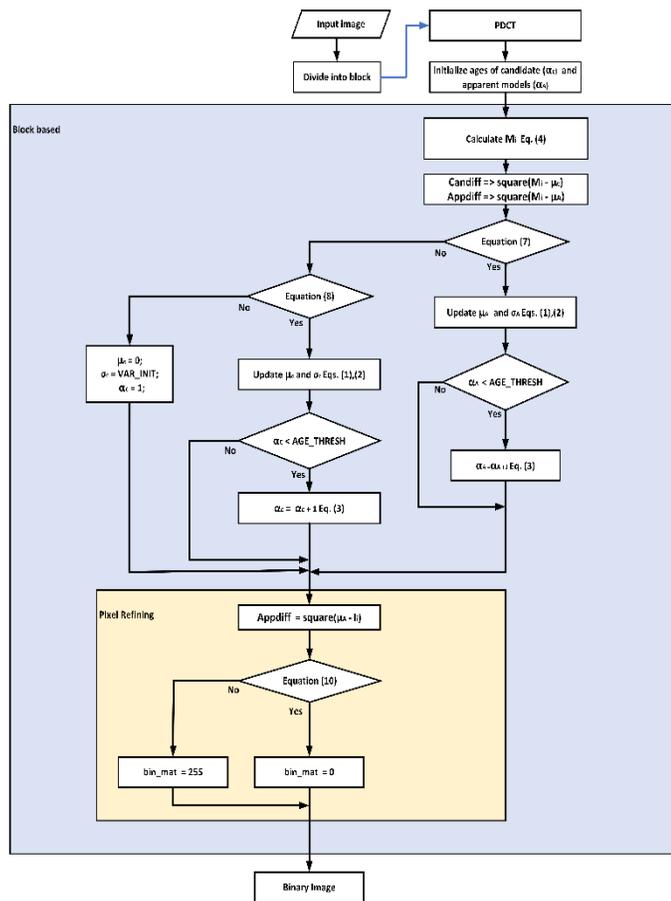


Fig. 3. A Flowchart for the DSGM Process.

$$(I_j^{(t)} - \mu_{A,i}^{(t)})^2 < \theta_d \sigma_{A,i}^{(t)} \quad (10)$$

where  $\theta_d$  is a threshold parameter. So, instead of the apparent background model learning the foreground data, the candidate background model learns it. Additionally, the correct background model will be chosen if the candidate background model's age is greater than the apparent background model's, where the models will be swapped. As a result, we don't have to be concerned about the model learning inaccurate foregrounds.

#### F. Computation Complexity

The quantity of elements processed in every frame determines the difficulty of the computation. We can only evaluate the computing complexity of one block because each frame is divided into equal-sized blocks of size  $8 \times 8$  pixels. Because each frame is broken into blocks of  $8 \times 8$  pixels of similar size, we may calculate the computing cost of a single block.

- For the CS process, we consider the original MoG [9], where each pixel is modelled by 3 Gaussians, which means that we need  $64 \times 3$  Gaussians per block.
- For CS-MoG [26], where each projection value requires three Gaussians, the number of Gaussians required per block is  $8 \times 3$  (a factor of 8 reduction).
- For our proposed method, each block is modelled by 2 Gaussian DM-SGM and we proceed over the 10 low-frequency DCT components, which require  $10 \times 2$  number of Gaussians for each block, a reduction by a factor of 9.6 and 1.2 per block w.r.t. the original MoG and CS-MoG, respectively. Experiments show that it is 2.5 times faster than CS-MOG in processing time.

#### G. Scene Reconstruction

When an image arrives at the sink node, it is superimposed on the previously received reference frame. Because the suggested technique only communicates a fraction of the entire image, the pixel coordinates at the MN node stay unchanged. This allows a portion of the transmitted image to be used to replace pixels in the reference image at the sink node more efficiently. The pixel values are, however, subject to channel distortion due to the transmission environment.

### IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, experimentation and performance evaluation are done to determine the relevance of our proposed method. The experimental dataset and setup are explained, then the qualitative analysis is shown to illustrate the performance of our system, and evaluation for quantitative and execution performance is done to test the accuracy and running time. In addition, the algorithm is also simulated in a sensor network environment using the Cooja Simulator of Contiki OS [36, 37] and realised in a real testbed using IOT-LAB [38].

#### A. Dataset and Setup

We will present the results of our compressed domain-based moving object detection technique on a standard

benchmark dataset, CDnet 2014<sup>1</sup> [39], to demonstrate its effectiveness. The CDnet 2014 data set is divided into 11 categories with different challenges, each of which contains four to six video sequences. Each video sequence consists of 600 to 7999 frames, with resolutions ranging from  $320 \times 240$  to  $720 \times 576$ . The simulations were run on an Intel Dual Core i7 3.6GHz processor with 8GB of RAM. The code is written in the C++ language. The total number of frame sequences in each dataset was averaged during the experiments.

#### B. Qualitative Analysis

Fig. 4 and 5 show the results of our moving object detection technique, Spectral Dual-Mode Background Subtraction (SDMBS). Fig. 4 exhibits performance for some of the representative frames from CDnet 's different categories to show performance against all the CDnet 2014 challenges. Fig. 4 and 5 demonstrate the ground truth and detected object discoveries from the original video frames. Comparing the resulting foreground mask to the relevant ground truth demonstrates the robustness of our suggested strategy for detecting moving objects across different categories.

Most of the CDnet 2014 challenges have excellent qualitative performance; nevertheless, the PTZ and camera jitter categories, as shown in Fig. 4, have poor qualitative performance. The worst performance Due to the zooming and moving features of this category, a compensation stage is required before the object detection stage to compensate for the frame movement. Because of the ghosting artefacts created in the videos in this category, the Intermittent Object Motion category is noisy. Background items moving away, abandoned objects, and objects stopping for a brief moment before moving away are the key features of this category. Shadow appearance in the shadow category affects the performance and the foregrounds are not detected completely.

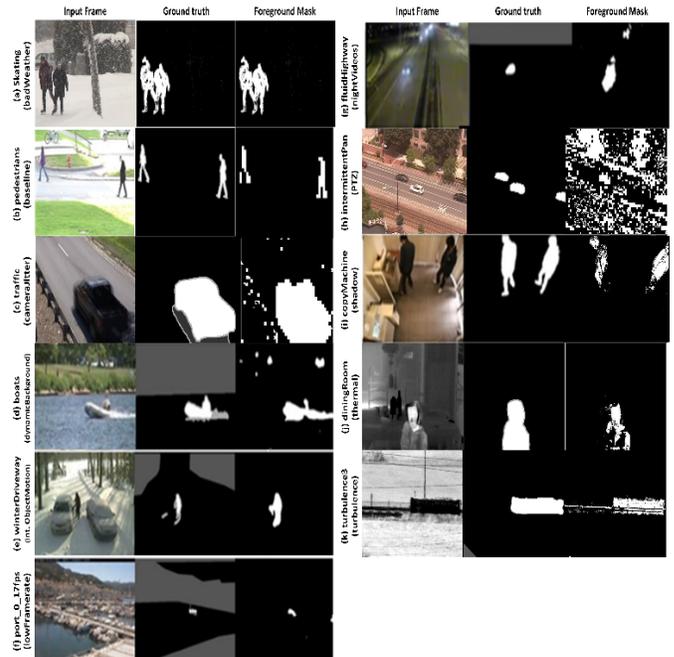


Fig. 4. Foreground Results of CDnet 2014 Dataset [39].

<sup>1</sup> <http://www.changedetection.net/>

When we compared our results to different existing methods published on the CDnet website [39], we identified MOG [9], KNN [40], ViBe [12], and SubS [41] as candidates. Thus, we compared our proposed compressed-based background subtraction SDMBS with recent and state-of-the-art methods [26,42], classical methods like [9,40], and fast methods like the ViBe [12] Background Subtraction Algorithm.

In [26], a block-based MOG is designed to be processed using the compressed sensing CS elements of the frame-blocks CS-MOG and is targeted at WVSN applications, whereas [42] is a background model update algorithm that uses an intermittent technique along with an adaptive block-learning algorithm.

The results of three video sequences, Highway (baseline), Fountain2 (dynamic background), and Snowfall (bad weather), are illustrated in Fig. 5. The original video frame for the three datasets and its corresponding groundtruth are shown in the top two rows. The results of MOG [9], Vibe [12], two current state-of-the-art techniques [26, 42], and SDMBS are shown in the next five rows (from top to bottom). In the last row of Fig. 5, we demonstrate a qualitative comparison of our proposed technique with other current methods, revealing that our method outperforms several of the existing systems. From the results, it is observed that our system accurately recognises foreground objects and has a considerably high resemblance to the ground truth when compared to other examined systems.

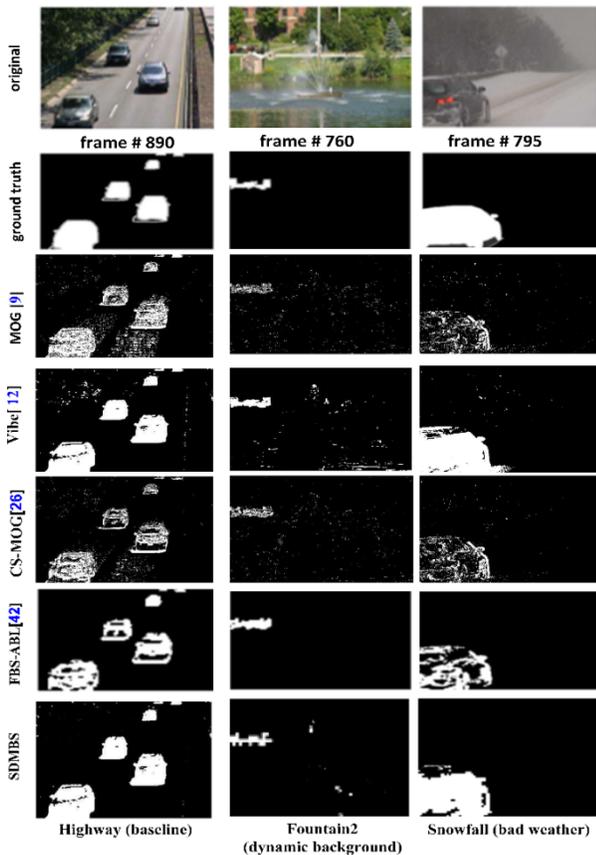


Fig. 5. Results with Highway (Baseline), Fountain2 (Dynamic Background), and Snowfall (Bad Weather) Videos Frames.

### C. Quantitative Analysis

In the quantitative evaluation, our method is compared to widely popular and state-of-the-art object detection algorithms for WVSN by conducting experimentation on the benchmark (CDnet 2014) dataset [39]. Several evaluation metrics are utilised to provide a credible measure of the outcome. Average recall (Re), precision (Pr), and F-measure (Fm) for all the video sequences in each category are listed in Table I. True positive (TP), false positive (FP), true negative (TN), and false-negative (FN) are the four types of pixel-based count metrics that can be created using the available ground truth data [39].

As the frequency of false negatives decreases, the value of Recall (Re) increases, which is used to measure the degree of completeness of the recognised foreground. Precision (Pr) is a metric measuring how accurate the identified foreground is, with a lower value when there are a lot of false positives. F-measure (Fm) is a metric for determining the balance of recall and precision with equal weights, implying that it is high only when both recall and precision are high. The three evaluation metrics, recall (Re), precision (Pr), and F-measure (Fm), are only considered to avoid redundancy.

TABLE I. EVALUATION METRICS

| Metrics                   | Description                      |
|---------------------------|----------------------------------|
| Recall (Re)               | $\frac{TP}{TP + FN}$             |
| Precision (Pr)            | $\frac{TP}{TP + FP}$             |
| F-Measure (Fm)            | $\frac{2(Pr \cdot Re)}{Pr + Re}$ |
| Specificity (SP)          | $\frac{TN}{TN + FP}$             |
| False Positive Rate (FPR) | $\frac{FP}{FP + TN}$             |
| False Negative Rate (FNR) | $\frac{FN}{TN + FP}$             |

The best and second-best performing approaches for each category, based on the average Fm for all the video sequences, are noted in red and bold in Tables II and III. When compared to classical methods, SDMBS may only be competitive in some areas, such as dynamic background, low frame rate, and bad weather. While there are approximate results for most categories with Subs [41] when SDMBS is ranked (2nd), this can be explained in terms of the design trade-off. While; when compared to state-of-the-art methods [26], we achieve a 15% increase in accuracy than CS-MOG [26] which is a compressed-based background subtraction applied for WVSN. In Fig. 7, the execution speed of SDMBS is compared to that of other methods at two resolution scales (320x240 and 640x480). For the two resolution scales, SDMBS excels in terms of speed. As seen in Fig. 7, SDMBS provides equivalent results to FBS-ABL [42], although it is more accurate, as seen in Fig. 6. When compared to other block-based techniques, this demonstrates SDMBS's effective design strategy.

TABLE II. COMPARISON ON THE FIRST SIX CATEGORIES OF CDNET 2014 DATASET.

| Category                   | Metrics | CDnet-14      | MOG[9] | KNN[40]       | ViBe[12] | SubS[41]      | CS-MOG [26] | FBS-ABL [42]  | SDMBS         |
|----------------------------|---------|---------------|--------|---------------|----------|---------------|-------------|---------------|---------------|
| Baseline                   | Re      | <b>0.9507</b> | 0.8180 | 0.7934        | 0.8204   | 0.9520        | 0.7557      | 0.8910        | 0.8775        |
|                            | Pr      | <b>0.9347</b> | 0.8461 | 0.9245        | 0.9288   | 0.9495        | 0.7942      | 0.8602        | 0.9481        |
|                            | Fm      | <b>0.9330</b> | 0.8245 | 0.8411        | 0.8700   | <b>0.9503</b> | 0.7745      | 0.8649        | <b>0.9114</b> |
| Dynamic background         | Re      | <b>0.8543</b> | 0.8344 | 0.8047        | 0.7222   | 0.7768        | 0.6534      | 0.7958        | 0.7359        |
|                            | Pr      | <b>0.8606</b> | 0.5989 | 0.6931        | 0.5346   | 0.8915        | 0.5262      | 0.7332        | 0.9604        |
|                            | Fm      | <b>0.8176</b> | 0.6330 | 0.6865        | 0.5652   | <b>0.8177</b> | 0.583       | 0.7424        | <b>0.8333</b> |
| Camera jitter              | Re      | <b>0.8159</b> | 0.7334 | 0.7351        | 0.7375   | 0.8243        | 0.6826      | 0.8046        | 0.3281        |
|                            | Pr      | <b>0.8359</b> | 0.5126 | 0.7018        | 0.4862   | 0.8115        | 0.4562      | 0.4656        | 0.5371        |
|                            | Fm      | <b>0.7806</b> | 0.5969 | <b>0.6894</b> | 0.5720   | <b>0.8152</b> | 0.5469      | 0.5298        | 0.4074        |
| Intermittent Object motion | Re      | <b>0.7231</b> | 0.5142 | 0.4617        | 0.5122   | 0.6578        | 0.4102      | 0.7861        | 0.5256        |
|                            | Pr      | <b>0.7888</b> | 0.6688 | 0.7121        | 0.6515   | 0.7957        | 0.6012      | 0.7943        | 0.7639        |
|                            | Fm      | <b>0.6795</b> | 0.5207 | 0.5026        | 0.5074   | <b>0.6569</b> | 0.48766     | <b>0.7232</b> | 0.6227        |
| Shadow                     | Re      | <b>0.9222</b> | 0.7960 | 0.7478        | 0.7833   | 0.9419        | 0.7462      | 0.9143        | ND            |
|                            | Pr      | <b>0.8551</b> | 0.7156 | 0.7788        | 0.8342   | 0.8646        | 0.6366      | 0.8569        | ND            |
|                            | Fm      | <b>0.8778</b> | 0.7370 | 0.7468        | 0.8032   | <b>0.8986</b> | 0.687       | <b>0.8671</b> | ND            |
| Thermal                    | Re      | <b>0.7727</b> | 0.5691 | 0.4817        | 0.5435   | 0.8161        | 0.5131      | 0.6394        | 0.7277        |
|                            | Pr      | <b>0.8795</b> | 0.8652 | 0.9186        | 0.9363   | 0.8328        | 0.8022      | 0.8002        | 0.8116        |
|                            | Fm      | <b>0.7962</b> | 0.6621 | 0.6046        | 0.6647   | <b>0.8171</b> | 0.6258      | 0.6619        | <b>0.7673</b> |

TABLE III. COMPARISON ON THE NEWER CATEGORIES OF CDNET 2014 DATASET

| Category           | Metrics | CDnet-14      | MOG[9] | KNN[40] | ViBe[12] | SubS[41]      | CS-MOG [26] | FBS-ABL [42]  | SDMBS         |
|--------------------|---------|---------------|--------|---------|----------|---------------|-------------|---------------|---------------|
| Low frame rate     | Re      | <b>0.7732</b> | 0.5823 | 0.6290  |          | 0.8537        | 0.5323      | 0.6616        | 0.5934        |
|                    | Pr      | <b>0.6894</b> | 0.6894 | 0.6865  |          | 0.6035        | 0.6394      | 0.7313        | 0.7398        |
|                    | Fm      | <b>0.6437</b> | 0.5373 | 0.5491  |          | <b>0.6445</b> | 0.5809      | 0.6328        | <b>0.6585</b> |
| Bad weather        | Re      | <b>0.7531</b> | 0.7181 | 0.6537  |          | 0.8213        | 0.6881      | 0.7449        | 0.7978        |
|                    | Pr      | <b>0.8960</b> | 0.7704 | 0.9114  |          | 0.9091        | 0.7354      | 0.8965        | 0.9385        |
|                    | Fm      | <b>0.8124</b> | 0.7380 | 0.7587  |          | <b>0.8619</b> | 0.7109      | 0.8106        | <b>0.8624</b> |
| Night videos       | Re      | <b>0.6107</b> | 0.5261 | 0.5413  |          | 0.6570        | 0.4761      | 0.7498        | 0.6892        |
|                    | Pr      | <b>0.5438</b> | 0.4128 | 0.4298  |          | 0.5359        | 0.3628      | 0.4957        | 0.4425        |
|                    | Fm      | <b>0.5154</b> | 0.4097 | 0.4200  |          | <b>0.5599</b> | 0.4117      | 0.5272        | <b>0.5386</b> |
| PTZ                | Re      | <b>0.7932</b> | 0.6475 | 0.6980  |          | 0.8306        | 0.5975      | 0.8357        | ND            |
|                    | Pr      | <b>0.3325</b> | 0.1185 | 0.1979  |          | 0.2840        | 0.1685      | 0.2290        | ND            |
|                    | Fm      | <b>0.3844</b> | 0.1522 | 0.2126  |          | <b>0.3476</b> | 0.2628      | <b>0.3267</b> | ND            |
| Turbulence         | Re      | <b>0.7391</b> | 0.7913 | 0.7682  |          | 0.8050        | 0.7413      | 0.9468        | 0.8023        |
|                    | Pr      | <b>0.7790</b> | 0.4293 | 0.5117  |          | 0.7814        | 0.3793      | 0.4936        | 0.5392        |
|                    | Fm      | <b>0.7145</b> | 0.4663 | 0.5198  |          | <b>0.7792</b> | 0.5018      | 0.5564        | <b>0.6448</b> |
| CDnet 2014 average | Re      | <b>0.7805</b> | 0.6845 | 0.6649  | 0.6865   | 0.8124        | 0.6178      | 0.7972        | 0.6752        |
|                    | Pr      | <b>0.7543</b> | 0.6025 | 0.6787  | 0.7286   | 0.7508        | 0.5547      | 0.6687        | 0.7423        |
|                    | Fm      | <b>0.7288</b> | 0.5707 | 0.5937  | 0.6637   | <b>0.7408</b> | 0.5612      | 0.6584        | <b>0.6940</b> |

Fig. 6 and 7 highlight the trade-off between detection performance and execution speed, and as can be seen, extensively adaptable approaches have fast/practical execution at the cost of diminished performance. We achieve a 2.3x

improvement in frame rate (FPS) over CS-MOG [26], a compressed-based background subtraction method used for WVSAN. This shows an efficient decrease in processing time.

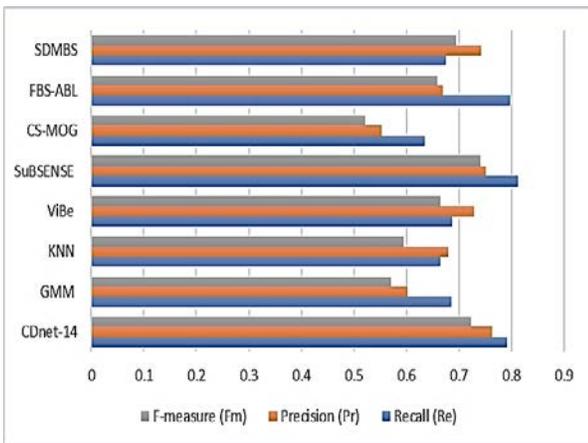


Fig. 6. Quantitative Analysis on CDnet Dataset.

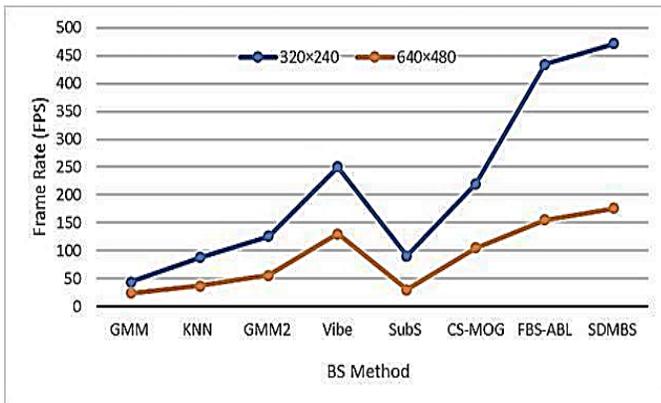


Fig. 7. Frame Rate for Different Techniques.

#### D. Sensor Network Simulation

This section illustrates the capability of the proposed system to work in WSN environments: first, simulation is carried out over Cooja of the Contiki OS Network Simulator [36], [37] to add the effect of lost packets and throughput. Second; the system is released on a real testbed using IOT-LAB [38]. Traffic trace files are used in the real testbed and simulated environment [15].

1) *Cooja simulation*: Four sensor nodes are installed. The sink is located at the left upper node (node 1) of the network area of 100 m × 100 m square grid. The destination node is located at (node 4). The simulation uses two datasets: pedestrians and PETS2006 (baseline) videos. The detection of moving objects is carried out at the host to select the blocks containing moving objects, and the blocks are then sent and routed through intermediate nodes to the base station. The received blocks at the destination are reconstructed to show the moving target, Fig. 8.

Fig. 9 shows the received image PSNR for two approaches: First, the full transmission of the image frame (Full Tx), while the second is our approach to transmitting the important portion of the image containing the moving object (Partial Tx). Although the proposed approach has a lower PSNR ratio than the full transmission approach, however, the average value is 27db. PSNR and energy are calculated using [15].

The proposed approach was compared to the direct approach for the energy consumption analysis. In the direct technique, a multi-hop transfer of an entire image to the sink node is used. On the aforementioned datasets, Fig. 10 depicts the energy using both methodologies. As can be observed, the node's energy usage has been significantly reduced. The two datasets show that the suggested approach can be employed in real-time moving object detection systems since a portion of the image data, including object information, is received at the sink node with an appropriate range of PSNR values and less energy.

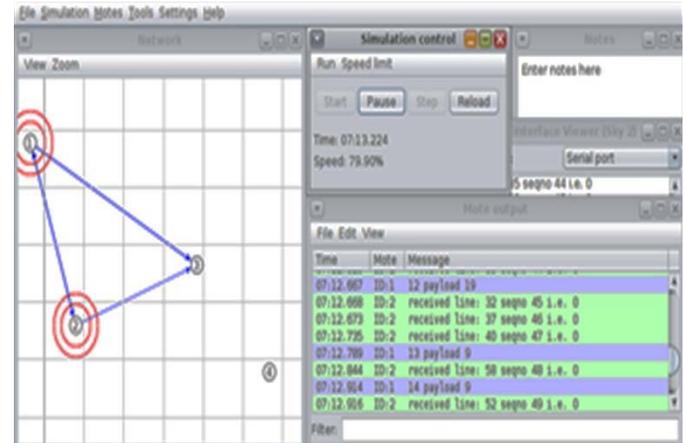


Fig. 8. Cooja Snapshot.

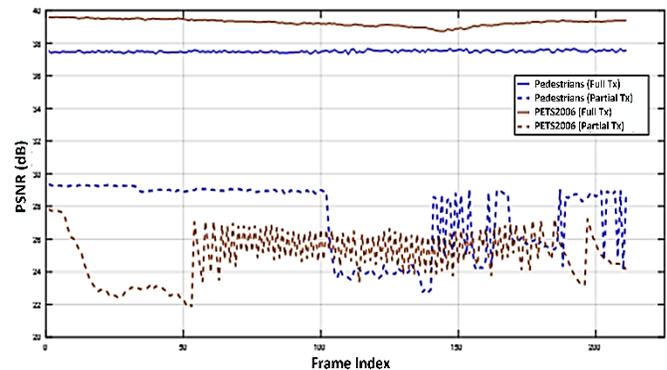


Fig. 9. PSNR for the Two Datasets.

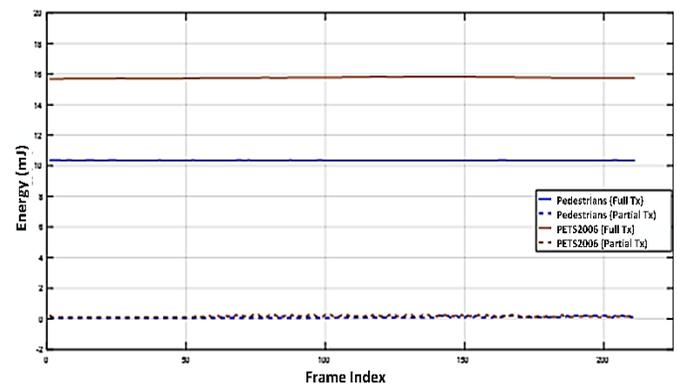


Fig. 10. Energy Consumption.

2) *IOT-LAB realization*: IoT-LAB<sup>2</sup> [38] is a large-scale WSN testbed that includes over 2000 wireless sensor nodes and a variety of processor architectures and wireless chips. IoT-LAB can be accessed through a web portal or by using the command-line tools. It allows users to retrieve experiment results and access serial ports on devices. Based on trace files as presented in [15], the IoT-LAB testbed M3 open nodes illustrated in Fig. 11(b) was employed in our experiments to replicate the intended object detection of the two datasets: pedestrians and PETS2006 (baseline). As shown in Fig. 11(a), the nodes m3-1, m3-10, m3-15, and m3-16 are used as senders, and m3-24 (blue circles) is used as a receiver to acquire varied loss rates as shown in Fig. 11(a). The sender (sender node) sends data packets according to the sender's trace file specifications (st-packet). The receiver (receiver node) maintains track of the packets it receives in a receiver trace file (rt-packet) as shown in Fig. 11. The sequence numbers of correctly received packets are received on the user's computer, which is used to reconstruct the video and calculate experiment metrics.

Fig. 12 shows the results of applying the proposed moving object detection technique in IOT-LAB to the two datasets: pedestrians on the first row and PETS2006 on the second row. The foreground blocks are transmitted and routed to the sink node. The sink node decompresses the received block and determines the moving object's location. For surveillance applications, object location is the most important piece of information that requires further analysis. The object ROI is transmitted to the sink node correctly with minimum network resources, memory, and bandwidth. The energy is minimised with an accepted PSNR.

## V. CONCLUSION

A background subtraction method that is both computationally efficient and accurate has been developed for object tracking across the limited resources of Wireless Visual Sensor Networks (WVSNS). To address the computation bottleneck of processing for constrained sensor networks, we use partial DCT to reduce the data dimensions while preserving the information content. In addition, energy-efficient block-based dual-mode SGM is utilised for foreground block detection, where the image frame is divided into blocks and only blocks containing foreground pixels are further processed for the refining stage. The foreground pixels are determined and the moving object is located. In contrast to standard Compress Sensing CS, which compresses the entire frame, the target region of interest ROI in our proposed method is compressed, communicated, and routed toward the sink node for further analysis. Our experimental results show that our method is as efficient as traditional algorithms. Moreover, it is up to three times faster than the state-of-the-art WVSNS object detection methods, and 15% more accurate. For embedded camera networks, we demonstrate that our suggested technique can accurately detect a moving object in real-time. We applied the proposed detection method in a WSN environment using Cooja of the Contiki OS Network Simulator. We verified that

the energy required for transmitting the detected object to the sink node in our proposed detection method is lower than that of comparable methods at acceptable PSNRs. Finally; the system is released on a real testbed using IOT-LAB using testbed M3 open nodes.

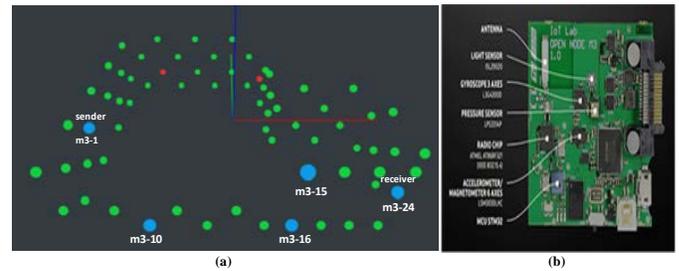


Fig. 11. IoT-LAB (a) Configuration, (b) Testbed M3 Open Nodes (ARM Cortex M3, 32-Bits MCU, and 802.15.4 PHY Standard).



Fig. 12. Object Detection Received at Destination Node Sink for the Two Datasets: Pedestrians (Upper Row) and PETS2006 (Lower Row).

## REFERENCES

- [1] T.C. H K Patil, Wireless Sensor Networks - an overview | ScienceDirect Topics, Am. Sci. Res. J. Eng. Technol. Sci. 64 (2017).<https://www.sciencedirect.com/topics/computer-science/wireless-sensor-networks>
- [2] I.F. Akyildiz, T. Melodia, K.R. Chowdury, Wireless multimedia sensor networks: A survey, *IEEE Wirel. Commun.* 14 (2007) 32–39. <https://doi.org/10.1109/MWC.2007.4407225>
- [3] Y. Ye, S. Ci, A.K. Katsaggelos, Y. Liu, Y. Qian, Wireless video surveillance: A survey, *IEEE Access.* 1 (2013) 646–660. <https://doi.org/10.1109/ACCESS.2013.2282613>
- [4] B. Ma, L. Huang, J. Shen, L. Shao, M.H. Yang, F. Porikli, Visual Tracking under Motion Blur, *IEEE Trans. Image Process.* 25 (2016) 5867–5876. <https://doi.org/10.1109/TIP.2016.2615812>
- [5] B. Garcia-Garcia, T. Bouwmans, A.J.R. Silva, Background subtraction in real applications: Challenges, current models and future directions, *Comput. Sci. Rev.* 35 (2020). <https://doi.org/10.1016/j.cosrev.2019.100204>
- [6] X. Li, M.K. Ng, X. Yuan, Median filtering-based methods for static background extraction from surveillance video, *Numer. Linear Algebr. with Appl.* 22 (2015) 845–865. <https://doi.org/10.1002/nla.1981>
- [7] S. Maity, A. Chakrabarti, D. Bhattacharjee, Block-Based Quantized Histogram (BBQH) for efficient background modeling and foreground extraction in video, in: 2017 Int. Conf. Data Manag. Anal. Innov. ICDMAI 2017, 2017: pp. 224–229. <https://doi.org/10.1109/ICDMAI.2017.8073514>
- [8] M. Boninsegna, A. Bozzoli, Tunable algorithm to update a reference image, *Signal Process. Image Commun.* 16 (2000) 353–

<sup>2</sup> <https://www.iot-lab.info/>

365. [https://doi.org/10.1016/S0923-5965\(99\)00063-6](https://doi.org/10.1016/S0923-5965(99)00063-6).
- [9] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real-time tracking, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2 (1999) 246–252. [https://doi.org/10.1109/cvpr.1999.784637\\_2](https://doi.org/10.1109/cvpr.1999.784637_2)
- [10] Elgammal, D. Harwood, L. Davis, Non-parametric model for background subtraction, in: *Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, 2000: pp. 751–767. [https://doi.org/10.1007/3-540-45053-x\\_48\\_2](https://doi.org/10.1007/3-540-45053-x_48_2)
- [11] K. Kim, T.H. Chalidabhongse, D. Harwood, L. Davis, Real-time foreground-background segmentation using codebook model, *Real-Time Imaging.* 11 (2005) 172–185. [https://doi.org/10.1016/j.rti.2004.12.004\\_2](https://doi.org/10.1016/j.rti.2004.12.004_2)
- [12] O. Barnich, M. Van Droogenbroeck, ViBe: A universal background subtraction algorithm for video sequences, *IEEE Trans. Image Process.* 20 (2011) 1709–1724. [https://doi.org/10.1109/TIP.2010.2101613\\_2](https://doi.org/10.1109/TIP.2010.2101613_2)
- [13] D.L. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory.* 52 (2006) 1289–1306. [https://doi.org/10.1109/TIT.2006.871582\\_2](https://doi.org/10.1109/TIT.2006.871582_2)
- [14] N. Ahmed, T. Natarajan, and K. R. Rao, “Discrete cosine transform,” *IEEE Transactions on Computers*, vol. 23, pp. 90–93, 1974.
- [15] M. Maimour, SenseVid: A traffic trace based tool for QoE Video transmission assessment dedicated to Wireless Video Sensor Networks, *Simul. Model. Pract. Theory.* 87 (2018) 120–137. [https://doi.org/10.1016/j.simpat.2018.06.006\\_2](https://doi.org/10.1016/j.simpat.2018.06.006_2)
- [16] R. Banerjee, S. Das Bit, Low-overhead video compression combining partial discrete cosine transform and compressed sensing in WMSNs, *Wirel. Networks.* 25 (2019) 5113–5135. [https://doi.org/10.1007/s11276-019-02119-y\\_2](https://doi.org/10.1007/s11276-019-02119-y_2)
- [17] K.M. Yi, K. Yun, S.W. Kim, H.J. Chang, H. Jeong, J.Y. Choi, Detection of moving objects with non-stationary cameras in 5.8ms: Bringing motion detection to your mobile device, in: *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, 2013: pp. 27–34. [https://doi.org/10.1109/CVPRW.2013.9\\_2](https://doi.org/10.1109/CVPRW.2013.9_2)
- [18] F. De La Torre, M.J. Black, A framework for robust subspace learning, *Int. J. Comput. Vis.* 54 (2003) 117–142. [https://doi.org/10.1023/A:1023709501986\\_2](https://doi.org/10.1023/A:1023709501986_2)
- [19] Xiaowei Zhou, Can Yang, Weichuan Yu "Moving Object Detection by Detecting Contiguous Outliers in the Low-Rank Representation." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35:597-610, 2013.
- [20] A. Zheng, T. Zou, Y. Zhao, B. Jiang, J. Tang, C. Li, Background subtraction with multi-scale structured low-rank and sparse factorization, *Neurocomputing.* 328 (2019) 113–121. [https://doi.org/10.1016/j.neucom.2018.02.101\\_2](https://doi.org/10.1016/j.neucom.2018.02.101_2)
- [21] J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, *IEEE Trans. Inf. Theory.* 53 (2007) 4655–4666. [https://doi.org/10.1109/TIT.2007.909108\\_2](https://doi.org/10.1109/TIT.2007.909108_2)
- [22] M. Lamarre, J.J. Clark, Background subtraction using competing models in the block-DCT domain, in *Proc. - Int. Conf. Pattern Recognit.*, 2002: pp. 299–302. [https://doi.org/10.1109/ICPR.2002.1044695\\_2](https://doi.org/10.1109/ICPR.2002.1044695_2)
- [23] Y.A. Ur Rehman, M. Tariq, T. Sato, A novel energy efficient object detection and image transmission approach for wireless multimedia sensor networks, *IEEE Sens. J.* 16 (2016). [https://doi.org/10.1109/JSEN.2016.2574989\\_2](https://doi.org/10.1109/JSEN.2016.2574989_2)
- [24] D.M. Pham, S.M. Aziz, Object extraction scheme and protocol for energy efficient image communication over wireless sensor networks, *Comput. Networks.* 57 (2013) 2949–2960. [https://doi.org/10.1016/j.comnet.2013.07.001\\_2](https://doi.org/10.1016/j.comnet.2013.07.001_2)
- [25] S.A. Nandhini, S. Radha, R. Kishore, Efficient compressed sensing based object detection system for video surveillance application in WMSN, *Multimed. Tools Appl.* 77 (2018) 1905–1925. [https://doi.org/10.1007/s11042-017-4345-2\\_2](https://doi.org/10.1007/s11042-017-4345-2_2)
- [26] Y. Shen, W. Hu, M. Yang, J. Liu, B. Wei, S. Lucey, C.T. Chou, Real-time and robust compressive background subtraction for embedded camera networks, *IEEE Trans. Mob. Comput.* 15 (2016) 406–418. [https://doi.org/10.1109/TMC.2015.2418775\\_2](https://doi.org/10.1109/TMC.2015.2418775_2)
- [27] A. Tulsyan, B. Huang, R.B. Gopaluni, J.F. Forbes, Performance assessment, diagnosis, and optimal selection of non-linear state filters, *J. Process Control.* 24 (2014) 460–478. [https://doi.org/10.1016/j.jprocont.2013.10.015\\_2](https://doi.org/10.1016/j.jprocont.2013.10.015_2)
- [28] M. Khare, R.K. Srivastava, A. Khare, Moving object segmentation in Daubechies complex wavelet domain, *Signal, Image Video Process.* 9 (2015) 635–650. [https://doi.org/10.1007/s11760-013-0496-4\\_2](https://doi.org/10.1007/s11760-013-0496-4_2)
- [29] S.S. Sengar, S. Mukhopadhyay, Moving object detection using statistical background subtraction in wavelet compressed domain, *Multimed. Tools Appl.* 79 (2020) 5919–5940. [https://doi.org/10.1007/s11042-019-08506-z\\_2](https://doi.org/10.1007/s11042-019-08506-z_2)
- [30] W. Wang, J. Yang, W. Gao, Modeling background and segmenting moving objects from compressed video, *IEEE Trans. Circuits Syst. Video Technol.* 18 (2008) 670–681. [https://doi.org/10.1109/TCSVT.2008.918800\\_2](https://doi.org/10.1109/TCSVT.2008.918800_2)
- [31] S. Popa, D. Crookes, P. Miller, Hardware acceleration of background modeling in the compressed domain, *IEEE Trans. Inf. Forensics Secur.* 8 (2013) 1562–1574. [https://doi.org/10.1109/TIFS.2013.2276753\\_2](https://doi.org/10.1109/TIFS.2013.2276753_2)
- [32] H. Ye, J. Pei, Infrared images target detection based on background modeling in the discrete cosine domain, in: 2018: p. 33. [https://doi.org/10.1117/12.2285785\\_2](https://doi.org/10.1117/12.2285785_2)
- [33] P. Zhang, X. Wang, X. Wang, C. Fei, Z. Guo, Infrared small target detection based on spatial-temporal enhancement using quaternion discrete cosine transform, *IEEE Access.* 7 (2019) 54712–54723. [https://doi.org/10.1109/ACCESS.2019.2912976\\_2](https://doi.org/10.1109/ACCESS.2019.2912976_2)
- [34] S. Manimozhi, S. Aasha Nandhini, S. Radha, Compressed Sensing based background subtraction for object detection in WSN, in: 2015 Int. Conf. Commun. Signal Process. ICCSP 2015, 2015: pp. 569–573. [https://doi.org/10.1109/ICCSP.2015.7322550\\_2](https://doi.org/10.1109/ICCSP.2015.7322550_2)
- [35] S.W. Kim, K. Yun, K.M. Yi, S.J. Kim, J.Y. Choi, Detection of moving objects with a moving camera using non-panoramic background model, *Mach. Vis. Appl.* 24 (2013) 1015–1028. [https://doi.org/10.1007/s00138-012-0448-y\\_2](https://doi.org/10.1007/s00138-012-0448-y_2)
- [36] A. Dunkels, B. Grönvall, T. Voigt, Contiki - A lightweight and flexible operating system for tiny networked sensors, in: *Proc. - Conf. Local Comput. Networks, LCN*, 2004: pp. 455–462. [https://doi.org/10.1109/LCN.2004.38\\_2](https://doi.org/10.1109/LCN.2004.38_2)
- [37] F. Österlind, A. Dunkels, J. Eriksson, N. Finne, T. Voigt, Cross-level sensor network simulation with COOJA, in: *Proc. - Conf. Local Comput. Networks, LCN*, 2006: pp. 641–648. [https://doi.org/10.1109/LCN.2006.322172\\_2](https://doi.org/10.1109/LCN.2006.322172_2)
- [38] C. Adjih, E. Baccelli, E. Fleury, G. Harter, N. Mitton, T. Noel, R. Pissard-Gibollet, F. Saint-Marcel, G. Schreiner, J. Vandaele, T. Watteyne, FIT IoT-LAB: A large scale open experimental IoT testbed, in: *IEEE World Forum Internet Things, WF-IoT 2015 - Proc.*, 2015: pp. 459–464. [https://doi.org/10.1109/WF-IoT.2015.7389098\\_2](https://doi.org/10.1109/WF-IoT.2015.7389098_2)
- [39] Y. Wang, P.M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, P. Ishwar, CDnet 2014: An expanded change detection benchmark dataset, in: *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.*, 2014: pp. 393–400. [https://doi.org/10.1109/CVPRW.2014.126\\_2](https://doi.org/10.1109/CVPRW.2014.126_2)
- [40] Z. Zivkovic, F. Van Der Heijden, Efficient adaptive density estimation per image pixel for the task of background subtraction, *Pattern Recognit. Lett.* 27 (2006) 773–780. [https://doi.org/10.1016/j.patrec.2005.11.005\\_2](https://doi.org/10.1016/j.patrec.2005.11.005_2)
- [41] P.L. St-Charles, G.A. Bilodeau, R. Bergevin, SuBSENSE: A universal change detection method with local adaptive sensitivity, *IEEE Trans. Image Process.* 24 (2015) 359–373. [https://doi.org/10.1109/TIP.2014.2378053\\_2](https://doi.org/10.1109/TIP.2014.2378053_2)
- [42] V.J. Montero, W.Y. Jung, Y.J. Jeong, Fast background subtraction with adaptive block learning using expectation value suitable for real-time moving object detection, *J. Real-Time Image Process.* 18 (2021) 967–981. [https://doi.org/10.1007/s11554-020-01058-8\\_2](https://doi.org/10.1007/s11554-020-01058-8_2)

# Enhancing the Security of Digital Image Encryption using Diagonalize Multidimensional Nonlinear Chaotic System

Mahmoud I. Moussa<sup>1</sup>

Computer Science Department  
Faculty of Computer and Artificial  
Intelligence  
Benha University, Egypt

Eman I. Abd El-Latif<sup>2</sup>

Mathematics Department  
Faculty of Science, Benha University  
Benha University, Egypt

Nawaz Majid<sup>3</sup>

Computer Science Department  
Faculty of Science, Northern Border  
University (NBU)  
KSA

**Abstract**—This paper describes a new efficient cryptosystem for the color image encryption technique, based on a combination of multidimensional proposed chaos systems. This chaos system consists of six bisections:  $T_1(x), T_2(x), T_2(y), T_3(x), T_3(y)$ , and  $T_3(z)$ . They induce three chaotic matrix keys and three chaotic vector keys. We use a multidimensional chaotic system together with an encryption algorithm to provide better security and wide key spaces. The proposed cryptosystem uses four levels of random pixel diffusions and permutations simultaneously and  $\omega$  - times interchange between rows and columns. The correlations between the RGB components of the plain image are reduced. The level of security, the computational complexity, the quality of decoding a decrypted image under closure threat is improved. The simulation results showed that the algorithm shows a high level of security, and the assurance that the image recovered at the receiving point is identified as the image at the transmission point.

**Keywords**—Chaotic system; encryption; decryption; image; algorithms; cryptosystem

## I. INTRODUCTION

The world is moving towards digitizing all types of data. The digital use of data is hourly increasing. Digital images are at the forefront of the most used data on the internet. With the increase in the technology of the World Wide Web and the growth of its uses, the sharing of data over the internet has diversified, and among this data is the exchange of digital images. In many cases, the images are sensitive, and they contain information that is not publishable, such as images related to the health conditions of individuals, as well as secret military locations related to the national security of countries. Since the areas of communication via the Internet are open and vulnerable to attacks on data, the use of encryption for such images has become very important and necessary. A color image holds many rows and columns of pixels; these pixels contain quantized values that represent the degree of the color at any point in the image. Thus, the color image defined as a matrix of value pixels, each containing three numerical RGB components to describe the color of a tiny area. Since sensitive images may invite attacks from anywhere around the world, image security is an important issue. The aim of image encryption is to transform a plain image to different cipher one that is difficult to read [1]. Chaos theory is one of the most

important security approaches used in encrypting images due to its capacity for mixing, sensitivity to initial conditions, control parameters, and completely random behavior. In 1989 [2], Matthews developed the first chaotic encryption technique. Many researchers are interested in presenting various algorithms to create a strong and robust encryption system for digital images, and some of them depend on the chaos system to induce random secret keys to maintain the confidentiality of images. In this paper, we used a multi-dimensional chaos system with many parameters to increase the key space. Since the proposed chaos system is developed and new, the chaotic and bifurcation behavior of it have been made and the values that enter the system into a chaos state have been determined. This paper presents a cryptosystem based on the proposed multi-dimensional chaotic maps and multiplexing frequent levels of shuffling, scramble, and pixel diffusion for the digital image component RGB. The practical results showed the intensity and resistance of the proposed algorithm compared to many related ones. The analysis proved a significant increase in the size of the encryption keys, which makes our algorithm outperform others against the brute-force attacks. His work introduced a low complexity algorithm for image cryptography using a multiple definition function for linear chaotic map denoted NPWLCM. Since then, many scientific papers have described image encryption based on chaotic systems. In this work, we describe a multi-chaotic system to encrypt the RGB components of color images simultaneously such that, the RGB components affect one another. The proposed six maps; ;  $T_1(x), T_2(x), T_2(y), T_3(x), T_3(y)$ , and  $T_3(z)$  increase conjugation of the items  $x_i^3, x_i^2, y_i^2, x_i y_i, y_i^3, z_i^2 x_i, \dots, z_i^3$ . Pixel Transform Table (PTT) procedure input these maps to output three  $M \times N$  matrix keys ( $M_r, M_b, M_g$ ) and three vector keys ( $V_r, V_b, V_g$ ) of length  $MN$ . The image's rows and columns are shuffled and scrambled using the vector keys. While the matrix keys are used to change the values of pixels four times to increase the complexity and thus the security of the cryptosystem. The remainder of this paper presents the related work in Section 2. Section 3 and section 4 explain in detail the proposed chaotic system and its chaotic behavior and bifurcation. In Section 5, we describe the encryption and decryption cryptosystem. Finally, in Section 6, we show the practical results and comparisons with another research.

## II. RELATED WORK

The encryption algorithms based on the chaos theory have been arisen as a power approach to increase the security over the last few decades. The vast majority image encryption algorithms have been introduced based on one- and two-dimensional chaotic maps. In 2017, Pak and Huang [3] described a new chaotic system created by composing the output of two existing one-dimensional chaotic maps. Their algorithm achieved a total shuffling based on a linear-nonlinear-linear structure. In 2007, Chong Fu et al. [4] used a three-dimensional Lorenz chaotic system to increase security and performance of image cryptography. In 2008, Xiangdong et al. [5] presented a chaotic shuffling algorithm using a sorting transformation of a chaotic sequence to obtain address codes for image transposition. Their algorithm avoided the drawbacks of image scrambling ways like rising of complexity and requiring understanding of probability distribution. In 2009, Juan et al. [6] described a three-dimensional image cryptography approach based on a discrete chaotic system with a security key induced from the initial conditions and parameters of the logistic system. In 2011 [7], researchers introduced an image encryption scheme using a Lorenz and Rossler chaotic system to obtain a large key space, improving security and complexity. In 2005, Zhang et al. [8] presented an algorithm based on a chaotic map and big size encryption key to encrypt the given image followed by a pixel shuffling process using an induced chaotic permutation matrix. In 2011, Keshari and Modani [9] presented an image cryptosystem based on two random maps; a chaotic map lattice to change pixel values by iterating the chaotic map for given initial conditions and Arnold's cat map to rearrange the pixel's position. The same year, Zhang, and Liu [10] introduced a novel image encryption algorithm using a skew tent chaotic system. In their work, they shuffle the order of the positions of all the pixels in the image. Their algorithm was based on permutation-diffusion architecture. In 2012, Khade and Narnaware [6] presented another method to encrypt a color image depending on 3D Logistic and Chebyshev maps. The proposed chaotic maps substituted the RGB components, generated a key, and scrambled the image pixels. The used technique depending on a logistic map used to depict a grey-scale image. A digital matrix approach is used where the three-dimensional matrix value is replaced according to the generated chaotic sequences, whereby the pixel replacement and mixing are achieved at the same time.

Color images are rich with information and have attracted wide attention. Each pixel contains numerical values of RGB components that determine density of RGB components in the color image. Many encryption algorithms have been described [11, 12, 13, 14, 15, 16, 17]. These algorithms are more vulnerable to attack because they neglect the correlations between RGB components. In 2012, Wang et al. [18] proposed a new algorithm using a three-dimensional matrix of the color image and a two-dimensional Lorenz and tent chaotic system at the same time to encrypt RGB components. Their algorithm has four phases. First, the three-dimensional matrix is converted to a two-dimensional matrix and the low-frequency wavelet coefficient is divided into overlapping blocks. Then, encryption is achieved by scrambling the pixel value diffusion based on a completely random chaotic sequence. In 2016,

Younes [19] published a useful survey of different techniques of image encryption, discussing several image encryption techniques from 2013 to 2015. In 2021, El Shafai et al. [20] introduced an encryption method to encrypt medical images depending on a piecewise linear chaotic map, and DNA encoding techniques. In 2018, Wu and Yang [21] introduced an algorithm depending on pixel diffusion and a DNA approach, which exploited the two-dimensional Hénon-Sine map to create a pixel permutation. In 2019, Wu et al. [22] used a nonlinear operation in cylindrical diffraction domain and compressed sensing to encrypt a multi-image based on an asymmetric approach. In 2013, Song et al. [23] defined a neighborhood nonlinear map and Coupled Map Lattices (CML) to describe a novel framework based on the Nonlinear Chaotic Algorithm (NCA) chaos and its spatiotemporal merits. In 2020, Yasser, et al. [13], induced an image encryption algorithm using a combination between Discrete Wavelet Transform (DWT) and a chaotic system to shuffle pixels and substitution operations.

## III. PROPOSED CHAOS SYSTEM

The proposed chaotic system consists of new 1D, 2D, and 3D equations derived from logistic map and cubic maps. The proposed system is a bijection and increases the quadratic and cubic coupling of the items  $y_i^2, x_i^2, x_i y_i, x_i^3, z_i^3$ , and it provides more security to the system. Six chaotic key maps will be derived within a limited range in (0,1).

$$\left\{ \begin{array}{l} \text{1D chaotic system} \\ x_{n+1} = r * x_n * (1 - x_n)(1 - \sin(x_n)) \\ \text{2D chaotic system} \\ x_{n+1} = r' * x_n * (1 - x_n)(1 - \sin(x_n)) + \gamma_1 y_n^2 \\ y_{n+1} = \mu y_n (y_n + 2)(1 - \sin(x_n)) + \gamma_2 (x_n y_n + x_n^2) \\ \text{3D chaotic system} \\ x_{n+1} = \lambda x_n (2 + x_n)(1 - \sin(x_n)) + \beta y_n^2 x_n + \alpha z_n^3 \\ y_{n+1} = \lambda y_n (2 + y_n)(1 - \sin(y_n)) + \beta z_n^2 y_n + \alpha z_n^3 \\ z_{n+1} = \lambda z_n (2 + z_n)(1 - \sin(z_n)) + \beta x_n^2 z_n + \alpha y_n^3 \end{array} \right\} \quad (1)$$

The control parameters beyond the range have no chaotic behavior. Fig. 1 shows that the 1D and 2D systems enter a chaotic condition and outputs a chaotic series in the region (0,1) subject to:  $5.77 < r', r < 12.97$ ,  $2.68 < \mu < 3.6$ ,  $0.039 < \gamma_1, \gamma_2 < 0.25$ . Fig. 2 shows that, the 3D system enters a chaotic condition in the region (0,1) subject to:  $3.49 < \lambda < 3.83$  and  $0.039 < \beta, \alpha < 0.043$ . The Bifurcation diagram of 1D, 2D, and 3D are shown in Fig. 3 and Fig. 4, respectively.

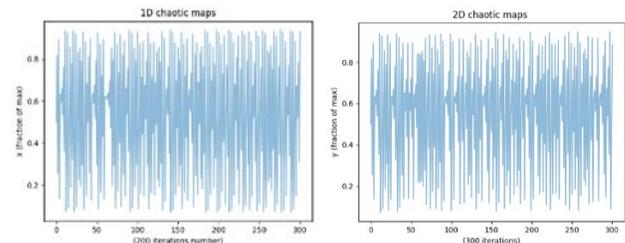


Fig. 1. The Behavior of the (1,2)-D chaotic Map in First 300 Iteration at  $r = 6.27$  in  $x$ - $y$  Plane.

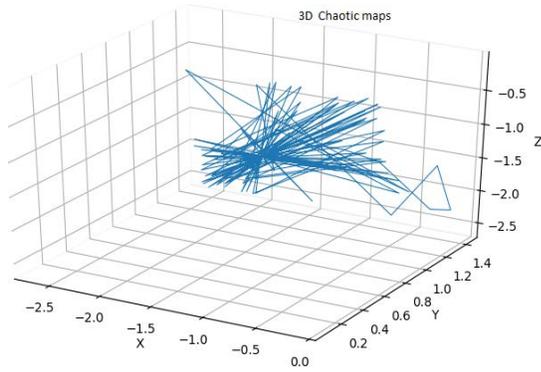


Fig. 2. The Behavior of the 3D Chaotic Map in First 1000 Iteration at  $\lambda = 3.81$ ,  $\beta = 0.41$ , and  $\alpha = 0.046$  in x-y-z- Space.

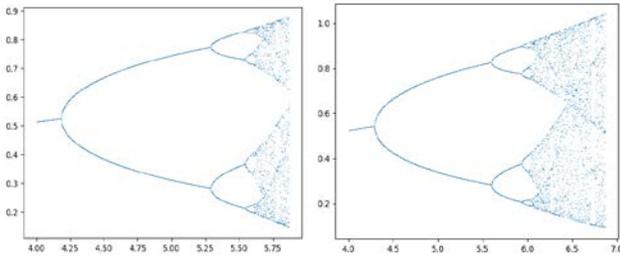


Fig. 3. Bifurcation Diagram of Sequences in 1D (2D) Chaotic System.

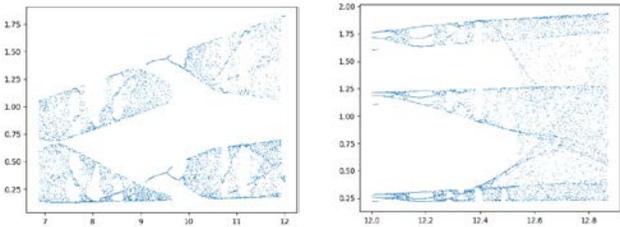


Fig. 4. Bifurcation Diagram of Sequences in 3D Chaotic System.

#### IV. DIAGONALIZING OF THE PROPOSED CHAOTIC MAP

For a given prime number,  $n$ , the proposed system in (1) can be represented as the set  $T$  of bijections as follows:

$$T = (T_1(x), T_2(x), T_3(x), T_2(y), T_3(y), T_3(z)), \quad (2)$$

where each bijection  $T_i(\dots) \in T$  can be defined as:

$$T_i \begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} \times \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} \bmod n \quad (3)$$

$$T_i \begin{bmatrix} x_{n+1} \\ y_{n+1} \\ z_{n+1} \end{bmatrix} = A \times \begin{bmatrix} x_n \\ y_n \\ z_n \end{bmatrix} \bmod n. \quad (4)$$

If the determinant  $|A|$  is not equal to zero and  $\gcd(|A|, n) = 1$  is satisfied, then the matrix  $A$  is invertible. The inverse of system (1) is:  $x_{n+1} = (A^{-1} \times x_n) \bmod n$ . Diagonalizing equations in (1) means finding a new  $x', y', z'$  system with no cross terms. Based on the Principal Axes theorem every quadratic form can be diagonalized. The diagonalization of the quadratic form 2D is  $(x'_{n+1})^T D x'_{n+1}$ ,  $(y'_{n+1})^T Q y'_{n+1}$ , where  $D$  and  $Q$  are  $2 \times 2$  matrices within four parameters for each. On the other side, diagonalization of the

3D form of the three variables  $x, y$ , and  $z$  increases the number of parameters up to nine [12].

#### V. PROPOSED SCHEME

##### A. Chaotic Key Generation

A color image  $P$  decomposes to three components  $RGB$ : red, blue, and gray with  $M \times N$  matrices each component with  $M$  rows and  $N$  columns of pixels. The proposed algorithm uses a novel scheme called Pixel Transform Table (PTT) to generate three  $M \times N$  matrix keys  $(M_r, M_b, M_g)$ , and three vector keys  $(V_x, V_y, V_z)$  of length  $MN$ . The chaotic maps  $T_1(x)$ ,  $T_2(x)$ , and  $T_2(y)$  generate three sequences of real numbers, they are converted to  $(M_r, M_b, M_g)$ , and the chaotic maps  $T_3(x)$ ,  $T_3(y)$ , and  $T_3(z)$  generate three other sequences which are transformed into  $(V_x, V_y, V_z)$ . The Pixel Transform Table (PTT) approach is shown below.

##### PTT ( $\rho \geq MN$ , jD map)

Input: The chaotic system (1)

Output: Random map  $T_j$

1. Initialize the chaotic parameters in the maps (1).
2. Iterate the system (1) to generate the sequences  $S_j(i)$  using jD map, where  $j = 1 \rightarrow 6$ , and  $i = 1, 2, \dots, \rho$ .
3.  $S_j(i) = \{S_1(i), S_2(i), \dots, S_6(i)\}$ .
4. For each  $j$ , generate the set of integer sequences  $S_j^l(\rho)$

from the set sequences  $S_j(i)$  as:

$$S_j^l(\rho) = \left[ (S_j(i) \times 10^{14}) \right] \bmod \rho$$

5. For each  $j = 1 \rightarrow 6$   
 $S^j(\rho) = \text{Sort}(S_j^l(\rho))$   
 find out the position of values  $S^j(\cdot)$  in  $S_j^l(\cdot)$ , then construct set of transfer  $T^j = \{t_1(i), t_2(i), \dots, t_6(i)\}$ , where the value  $S_j^l(t_j(i)) = S^j[i]$ ,  $i = 1, 2, \dots, \rho$ .
6. The sequences  $(t_1(i), t_2(i), t_3(i))$  are moved to three  $M \times N$  matrix keys  $(M_r, M_b, M_g)$ , and the sequences  $(t_4(i), t_5(i), t_6(i))$  are converted to the three vector keys  $(V_x, V_y, V_z)$  of length  $MN$ , respectively.

##### B. Image Encryption Algorithm (IEA)

The encryption is an operator that transforms the plain image  $P$  to an unknown cipher image  $P_5$ . IEA algorithm involves three phases as follows:

$$IEA: \text{Plain Image } (P) \rightarrow \text{Cipher Image } P_5$$

**Phase One:** We describe a new and double secure block shuffling mechanism together with pixels values change. The primary effect of PPT in this process is initially changing randomly the pixel values of  $RGB$  called 1<sup>st</sup> diffusion level as follows:

$$R^r = M_r R (M_r)^T \bmod 256$$

$$B^b = M_b B (M_b)^T \bmod 256$$

$$G^g = M_g G (M_g)^T \text{ mod } 256$$

where  $R^r$ ,  $B^b$ , and  $G^g$  are the producing matrices that contains the chaotic values. In each matrix ( $R^r, B^b, G^g$ ), we swap randomly the rows with an odd index ( $2k + 1$ ) with the rows that have an even index ( $2k$ ), and randomly interchange the columns with an odd index ( $2l + 1$ ) with the columns of an even index ( $2l$ ), as shown in Fig. 5. We repeat the swap process  $\omega$ -times and get ( $R^r(\omega), B^b(\omega), G^g(\omega)$ ).

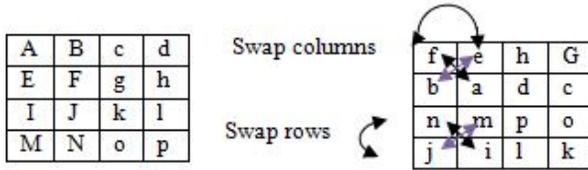


Fig. 5. One-Time Swapping Rows and Columns of the Matrix ( $R^r, B^b, G^g$ ).

Phase Two: In this phase, the pixel positions were scrambled in columns and rows by using two vector keys ( $V_x, V_y$ ). We apply 2<sup>nd</sup> diffusion level by dividing the components  $R^r(\omega), B^b(\omega),$  and  $G^g(\omega)$  values by their size  $MN$  to result the pixels values, i.e.

$$R'(\omega) = \left\lfloor \frac{B^r(\omega)}{MN} \right\rfloor \text{ mod } 256$$

$$B'(\omega) = \left\lfloor \frac{B^b(\omega)}{MN} \right\rfloor \text{ mod } 256$$

$$G'(\omega) = \left\lfloor \frac{G^g(\omega)}{MN} \right\rfloor \text{ mod } 256$$

Combine the matrices  $R'(\omega), B'(\omega),$  and  $G'(\omega)$  horizontally to obtain the  $M \times 3N$  matrix  $P_1$ , and generate a sequence of  $3MN$  numbers  $Y_1 = y_1, \dots, y_{3NM}$  from the rows of  $P_1$ . Permute the row  $Y_1 = y_1, \dots, y_{3NM}$  by the vector key  $V_x$ . We get the scramble vector  $Y'_1 = y'_1, \dots, y'_{3NM}$ . Reshape the vector  $Y'_1$  into three  $M \times N$  matrices;  $Ry(\omega), By(\omega),$  and  $Gy(\omega)$ . This process is repeated again as follows: Combine the matrices  $Ry(\omega), By(\omega),$  and  $Gy(\omega)$  vertically to obtain the  $3M \times N$  matrix  $P_2$ , and generate a sequence of  $3MN$  numbers  $Z_1 = z_1, \dots, z_{3NM}$  from the columns of  $P_2$ . Permute the sequence  $Z_1 = z_1, \dots, z_{3NM}$  by the vector key  $V_y$ . We get the scramble vector  $Z'_1 = z_1, \dots, z_{3NM}$ . Reshape the vector  $Z'_1$  into three  $M \times N$  matrices;  $Rz(\omega), Bz(\omega),$  and  $Gz(\omega)$ . Combine these matrices horizontally, we get a new  $M \times 3N$  denoted  $P_3$ .

Phase Three: Let  $D_{now}$  be the current ciphered pixel value after the current diffusion,  $P_{now}$  the present plain pixel value,  $D_{pre}$  the old cipher pixel value after the previous diffusion,

and  $P_{pre}$  the old plain value. Their initial values in  $Rz(\omega), Bz(\omega),$  and  $Gz(\omega)$  are equal to zero. Keep track of rows in  $P_3$  to set three vectors  $V(Rz(\omega)), V(Bz(\omega)),$  and  $V(Gz(\omega))$  each with length  $MN$ . We apply 3<sup>rd</sup> level of pixel diffusion on these vectors using the vector product with the key vector  $V_z$  as in the following:

$$\begin{cases} Zz^r = (V(Rz(\omega)) \times V_g) \text{ mod } 256 \\ Zz^b = (V(Bz(\omega)) \times V_g) \text{ mod } 256 \\ Zz^g = (V(Gz(\omega)) \times V_g) \text{ mod } 256 \end{cases} \quad (5)$$

where  $Zz^r, Zz^b,$  and  $Zz^g$  are the producing vectors that contains the chaotic values of  $V_z$ . We apply 4<sup>th</sup> and final level of pixel diffusion as shown in the next procedure where  $[(:)]$  refers to the components  $Zz^r, Zz^b,$  and  $Zz^g$  while  $[:]$  refers to a random value.

| <b>4<sup>th</sup> random Diffusion</b> ( $[(:)], [:]$ ) |                                                                                                                                                                                                                                                 |
|---------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Input:                                                  | $V_z = (Zz^r, Zz^b, Zz^g)$                                                                                                                                                                                                                      |
| Output:                                                 | Vector ( $D_{now}(Zz^r), D_{now}(Zz^b), D_{now}(Zz^g)$ )                                                                                                                                                                                        |
| 1.                                                      | Generate three random values follows:<br>$\epsilon_{1l} = \text{rand}() \% 3$<br>$\epsilon_{2l} = \text{choice a random value from } \{V_z\} \text{ mod } 256$<br>$\epsilon_{3l} = \text{choice a random value from } \{V_z\} \text{ mod } 256$ |
| 2.                                                      | For $\omega$ in range $MN$ do                                                                                                                                                                                                                   |
| 3.                                                      | If $\epsilon_{1l}=0$ then<br>$D_{now}(Zz^r) = (\epsilon_{2l} \cdot P_{now}(Zz^r) + \epsilon_{3l} \cdot (D_{pre}Zz^r \times P_{pre}Zz^r)) \text{ mo } 256.$ (6)                                                                                  |
| 4.                                                      | Else If $\epsilon_{1l}=1$<br>$D_{now}(Zz^b) = (\epsilon_{2l} \cdot P_{now}(Zz^b) + \epsilon_{3l} \cdot (D_{pre}Zz^b \times P_{pre}Zz^b)) \text{ mo } 256.$ (7)                                                                                  |
| 5.                                                      | Else If $\epsilon_{1l}=2$<br>$D_{now}(Zz^g) = (\epsilon_{2l} \cdot P_{now}(Zz^g) + \epsilon_{3l} \cdot (D_{pre}Zz^g \times P_{pre}Zz^g)) \text{ mo } 256.$ (8)                                                                                  |
| 6.                                                      | End For.                                                                                                                                                                                                                                        |

Reshape  $D_{now}(Zz^r), D_{now}(Zz^b),$  and  $D_{now}(Zz^g)$  into three  $M \times N$  matrices;  $Rz(\omega), Bz(\omega),$  and  $Gz(\omega)$ . We get the components of encrypted image  $P_5$  of size  $M \times N$ .

The flowchart of the encryption process with the induced 6 keys is shown in Fig. 6. The flowchart of the decryption algorithm and the 6 keys is shown in Fig. 7.

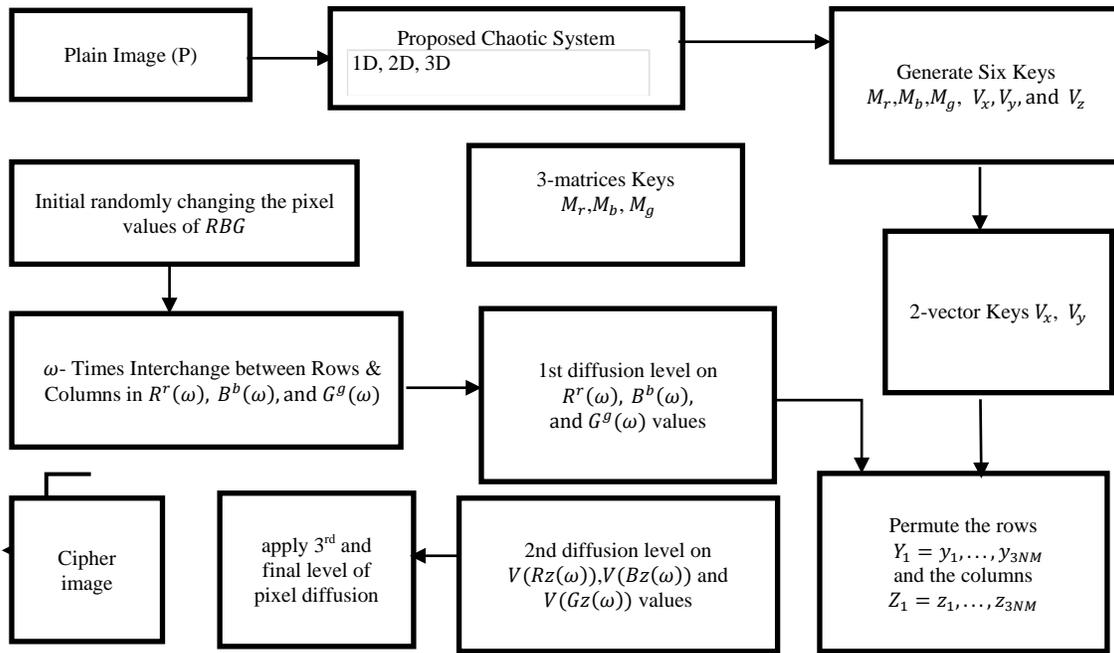


Fig. 6. The Outline of Image Encryption Algorithm IEA.

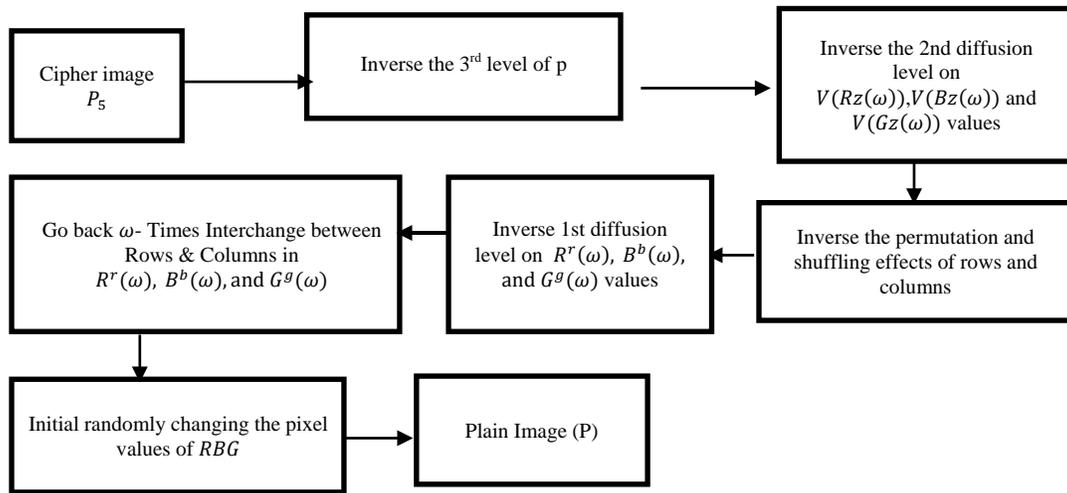


Fig. 7. The Outline of Image Decryption Algorithm IDA.

### C. Image Decryption Algorithm (IDA)

The image decryption algorithm (IDA) is like that of the image encryption pseudocode but in the inverse direction.

Step 1: All random values used in IEA are used in IDA.  
Step 2: Inverse 4<sup>th</sup> random Diffusion ( $[(:)]$ ,  $[:, ]$ ). Obtain  $Zz^r$ ,  $Zz^b$ , and  $Zz^g$  from  $D_{now}(Zz^r)$ ,  $D_{now}(Zz^b)$ , and  $D_{now}(Zz^g)$ . The inverse of the equations (6), (7), and (8) are the equations (9), (10), and (11) respectively.

$$Zz^r = \left( \frac{D_{now}(Zz^r) - \varepsilon_{3l}(D_{pre}Zz^r \times P_{pre}Zz^r)}{\varepsilon_{2l}} \right) \bmod 256 \quad (9)$$

$$Zz^b = \left( \frac{D_{now}(Zz^b) - \varepsilon_{3l}(D_{pre}Zz^b \times P_{pre}Zz^b)}{\varepsilon_{2l}} \right) \bmod 256 \quad (10)$$

$$Zz^g = \left( \frac{D_{now}(Zz^g) - \varepsilon_{3l}(D_{pre}Zz^g \times P_{pre}Zz^g)}{\varepsilon_{2l}} \right) \bmod 256 \quad (11)$$

Step 3: Inverse 3<sup>rd</sup> level of pixel diffusion. Obtain  $V(Rz(\omega))$ ,  $V(Bz(\omega))$  and  $V(Gz(\omega))$  from  $Zz^r$ ,  $Zz^b$ , and  $Zz^g$ , respectively. The inverse of the equations in (5) are the equations in (12).

$$\begin{cases} V(Rz(\omega)) = \left( \frac{Zz^r \times Vg}{Vg \cdot Vg} \right) \bmod 256 + \tau Vg \\ V(Bz(\omega)) = \left( \frac{Zz^b \times Vg}{Vg \cdot Vg} \right) \bmod 256 + \tau Vg \\ V(Gz(\omega)) = \left( \frac{Zz^g \times Vg}{Vg \cdot Vg} \right) \bmod 256 + \tau Vg \end{cases} \quad (12)$$

Step 4: Delete the effect of rows and columns scramble by using the inverse vectors  $(V_x^{-1}, V_y^{-1})$ .

Step 5: Inverse 2<sup>nd</sup> random Diffusion as follows:

$$B^r(\omega) = [MN \cdot R^l(\omega)] \bmod 256$$

$$B^b(\omega) = [MN \cdot B'(\omega)] \bmod 256 \quad (13)$$

$$G^g(\omega) = [MN \cdot G'(\omega)] \bmod 256$$

Step 6: Come back to the reverse paths to delete the effect of randomly interchange the columns rows.

Step 7: Inverse 1<sup>st</sup> random Diffusion as follows:

$$R = (M_r)^{-1} R^r ((M_r)^T)^{-1} \bmod 256$$

$$B = (M_b)^{-1} B^b ((M_b)^T)^{-1} \bmod 256 \quad (14)$$

$$G = (M_g)^{-1} G^g ((M_g)^T)^{-1} \bmod 256$$

These seven steps recover the plain image  $P$ . Mathematically, the image decryption algorithm is represented as in the next formula:

IDA: Cipher Image  $P_5 \rightarrow$  Plain Image ( $P$ )

#### D. The Computational Complicity

In this section, we discuss the running time of the steps in IEA. The running time of the PPT subroutine is  $O(\rho = MN + \epsilon)$  for small constant  $\epsilon$ , and the running time of the steps 1-7 is a linear function of the size of the image, i.e., the running time is  $O(MN)$ . The experimental process demonstrates that the consumption time of our approach IEA/IDA is practicable. The processor Intel(R) Core (TM) i7-8550U CPU @ 1.80 GHz 2.00 GHz and 8GB RAM is used in IEA/IDA is practicable. The processor Intel(R) Core (TM) i7-8550U CPU @ 1.80 GHz 2.00 GHz and 8GB RAM is used in encryption/decryption on 256 image  $P/P_5$  of size  $256 \times 256$ . The average consumption time is less than 4.01ms. Our running time is close to the running time in [24].

## VI. EXPERIMENTAL RESULTS AND HISTOGRAM ANALYSIS

The system is implemented by anaconda 4.8.3. Python is the programming language that we have picked for development. The implementation is examined in Windows 10 64-bit O.S. with an Intel(R) Core (TM) i7-8550U CPU @ 1.80GHz 2.00 GHz and 8GB RAM. Take the initial parameters and values: 1D ( $r = 5.78, x_0 = 1.2 \times 10^{18}$ ), 2D ( $r' = 6.37, \mu = 3.33, \gamma_1 = 0.17, \gamma_2 = 0.14, x_0 = 2.3 \times 10^{18}, y_0 = 1.2 \times 10^{18}$ ), 3D ( $\lambda = 3.66, \beta = 0.041, \alpha = 0.039, x_0 = 3.4 \times 10^{18}, y_0 = 4.5 \times 10^{18}, z_0 = 5.6 \times 10^{18}$ ), to encrypt "Baboon" and "Lena" images of size  $256 \times 256$  and their RGB components as shown in Fig. 8(a-d). The histograms of the "baboon", "Lena" and their RGB components before encryption are shown in Fig. 9(a-c); the histograms illustrate how pixels in the plain images "baboon" or "Lena" are correlated to the pixels at each color density level.

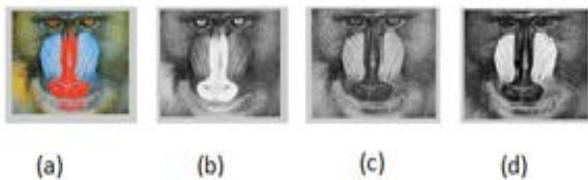


Fig. 8. (a-d) shows "Baboon" and "Lena" Color Images and the RGB Components of their Respective Color Images before the Encryption Process.

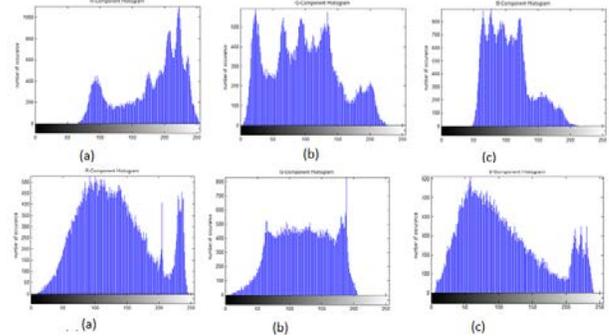


Fig. 9. Histograms of the RGB Components of "Baboon" and "Lena" before Encryption respectively.

Fig. 10 (a-d) shows the "baboon" and "Lena" encrypted images and their respective encrypted RGB components. The histograms of the RGB components of "baboon", and "Lena" after the encryption are shown below in Fig. 11 (a-c). The histograms illustrate how pixels in the ciphered RGB are uniformly correlated to the pixels at each color density level.

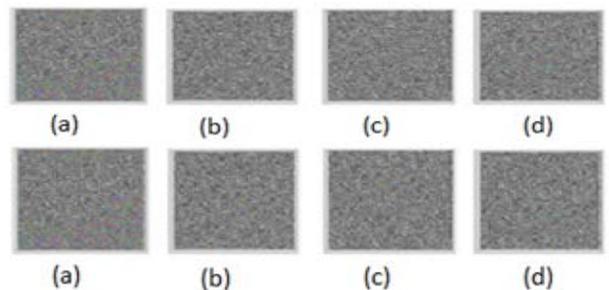


Fig. 10. (a-d) Shows "Baboon" and "Lena" Cipher images and their Ciphered RGB Components.

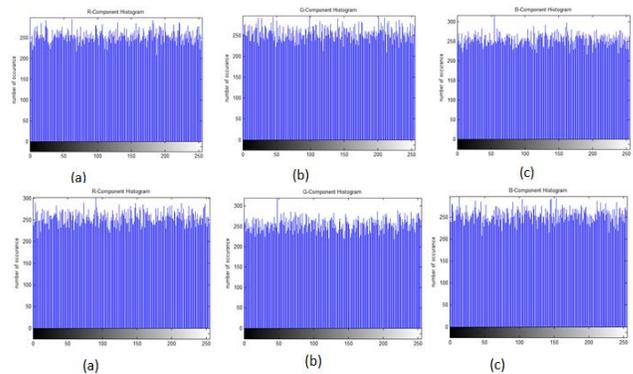


Fig. 11. (a-c): Histograms of RGB Component of "Baboon" and "Lena" after Encryption.

VII. SECURITY ANALYSIS

A. The Key Size

The total number of unique keys used in the encryption must be big enough to make brute-force attacks ineffective. Our algorithm employs six values;  $x_0^{1D}, x_0^{2D}, y_0^{2D}, x_0^{3D}, y_0^{3D}, z_0^{3D}$  and the eight parameters of  $\mu, r, r', \gamma_1, \gamma_2, \lambda, \beta, \alpha$ , as secret keys. Wang and Teng [25] proved that if the precision is  $10^{-17}$ , the keys  $K_{x_0^{1D}} = K_{x_0^{2D}} = K_{y_0^{2D}} = K_{x_0^{3D}} = K_{y_0^{3D}} = K_{z_0^{3D}} = 10^{17}$ ,  $K_\mu = K_r = K_{r'}, K_{\gamma_1} = K_{\gamma_2} = K_\lambda = K_\beta = K_\alpha = 0.5 \times 10^{17}$ . Let the plain color images have size  $256 \times 256$ . The number of iterations over six maps  $I_0$  is  $6 \times (3 \times M \times N) = 6 \times (3 \times 256 \times 256) \approx 2^{20} \approx 10^7$ . The total key space reaches to  $\approx 1.953 \times 10^7 \times 10^{235} = 1.953 \times 10^{242}$ . Our key space is larger than  $2^{138}, 2^{58}, 10^{140}, 2^{256}, 10^{79}, 4.2 \times 10^{122}, 10^{60}$ , and  $10^{112}$  [4, 13, 6, 9, 18, 26, 27, 28] respectively. It is greater than  $2^{448} = 7.8 \times 10^{134}$ , the maximum key space mentioned in the survey paper [14]. The total numbers of keys within the diagonalization form (1) reach 24 initial values and parameters. Our key increases it up to  $10^{415}$ . The proposed algorithms describe a sufficiently enough key space to withstand brute-force assaults.

B. The Sensitivity Analysis of the Secret Keys

Little differences between keys yield different cipher images. When the image is decrypted, using a wrong key induces another image. Fig. 12 displays the decrypted image of the Baboon with the proper key of  $\lambda=3.66$ . On the other side, Fig. 13 illustrates the decryption of the Baboon image with the incorrect encryption key  $\lambda=3.660000000000000001$ . It was successful in making the algorithm sensitive to the key. A small modification in the key will result in an entirely different decryption result, and the attacker won't be able to get to the right plain image.



Fig. 12. Result of Correct Parameters used to Decrypt the Baboon Image and its R, G and B Components.

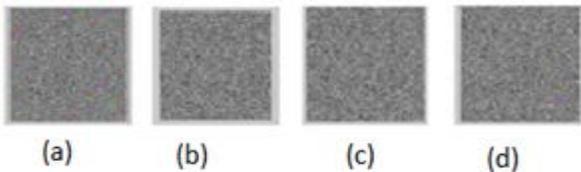


Fig. 13. Result of Wrong Parameters used to Decrypt the Baboon Image and its R, G and B Components.

C. Adjacent Pixels Correlation Analysis

The correlation between pixels is assessed by the degree of pixel association. In general, the stronger the correlation between nearby pixels in the ciphered image, the poorer the encryption algorithm's performance will be, and vice versa. The correlation in vertical, horizontal, and diagonal directions

between 3000 randomly selected nearby pixels is calculated as follows:

$$E(x) = \frac{1}{N} \sum_{i=1}^N (x_i) \tag{15}$$

$$E(y) = \frac{1}{N} \sum_{i=1}^N (y_i) \tag{16}$$

$$D(x) = \sum_{i=1}^N (x_i - E(x))^2 \tag{17}$$

$$D(y) = \sum_{i=1}^N (y_i - E(y))^2 \tag{18}$$

$$cov(x, y) = \frac{1}{N} \sum_{i=1}^N (x_i - E(x))(y_i - E(y)) \tag{19}$$

$$r_{xy} = \frac{cov(x, y)}{\sqrt{D(x)}\sqrt{D(y)}} \tag{20}$$

The values of two neighboring pixels,  $x$  and  $y$ , in the Horizontal direction H/D, Vertical Direction V/D and Diagonal direction D/D are shown in Tables I to III and Fig. 14. They exhibit a high level of concentration and a tendency of one in two neighboring pixels in the plain image and a correlation extremely near to 0 in the ciphered image, which means that adjacent pixels in the ciphered image are random, and the encryption impacts resist statistical attack.

TABLE I. TWO ADJACENT PIXEL CORRELATION VALUES FOR PLAIN AND CIPHERED IMAGES (V/D)

| Image  | V/D | Plain image | Cipher Image  |          |
|--------|-----|-------------|---------------|----------|
|        |     |             | New Algorithm | Ref [28] |
| Lena   | R   | 0.9238      | -0.0022       | -0.0016  |
|        | G   | 0.9479      | 0.0026        | -0.0011  |
|        | B   | 0.8785      | -0.00103      | -0.0013  |
| Baboon | R   | 0.9527      | -0.0021       | 0.0002   |
|        | G   | 0.9283      | -0.0047       | 0.0001   |
|        | B   | 0.9563      | 0.00201       | 0.0004   |

TABLE II. TWO ADJACENT PIXEL CORRELATION VALUES FOR PLAIN AND CIPHERED IMAGES (H/D)

| Image  | H/D | Plain image | Cipher Image  |          |
|--------|-----|-------------|---------------|----------|
|        |     |             | New Algorithm | Ref [27] |
| Lena   | R   | 0.9783      | -0.0017       | -0.00092 |
|        | G   | 0.9795      | 0.0034        | -0.0038  |
|        | B   | 0.9594      | -0.00063      | -0.0020  |
| Baboon | R   | 0.9413      | -0.0019       | 0.0062   |
|        | G   | 0.8796      | -0.0056       | -0.0060  |
|        | B   | 0.9164      | 0.0018        | 0.0077   |

TABLE III. TWO ADJACENT PIXEL CORRELATION VALUES FOR PLAIN AND CIPHERED IMAGES (D/D)

| Image  | Component D/D | Plain image | Cipher Image  |            |
|--------|---------------|-------------|---------------|------------|
|        |               |             | New Algorithm | Ref [10]   |
| Lena   | R             | 0.9685      | -0.0011       | -0.0008482 |
|        | G             | 0.9574      | 0.0021        | -0.0008482 |
|        | B             | 0.8994      | -0.00033      | -0.0008482 |
| Baboon | R             | 0.6471      | -0.0021       | 0.00370914 |
|        | G             | 0.9567      | -0.0036       | 0.00370914 |
|        | B             | 0.9355      | 0.0015        | 0.00370914 |

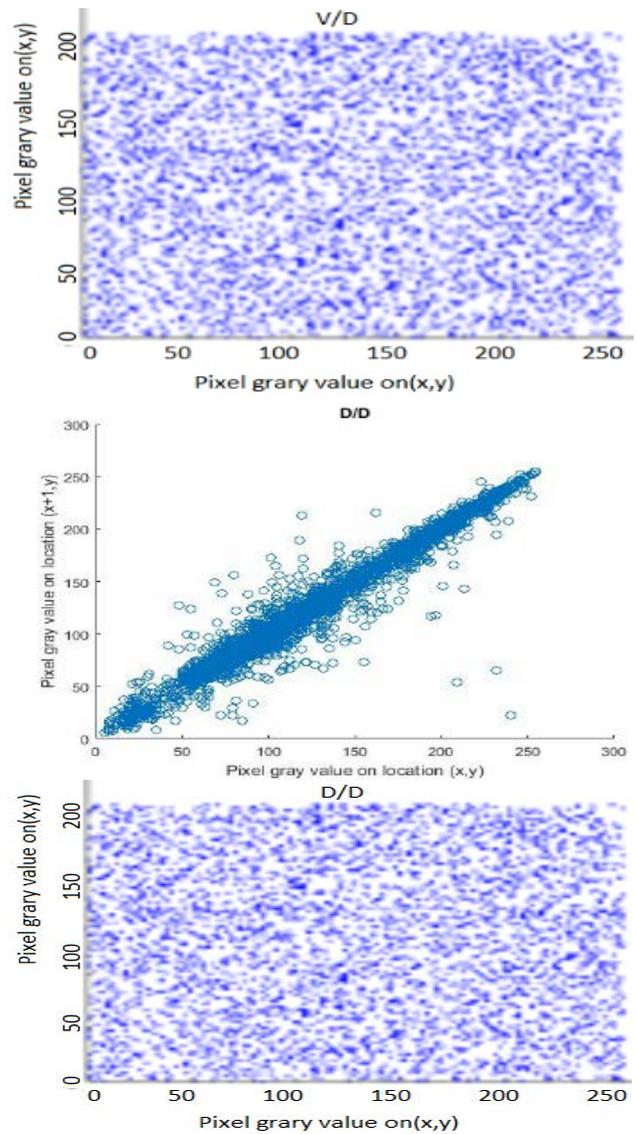
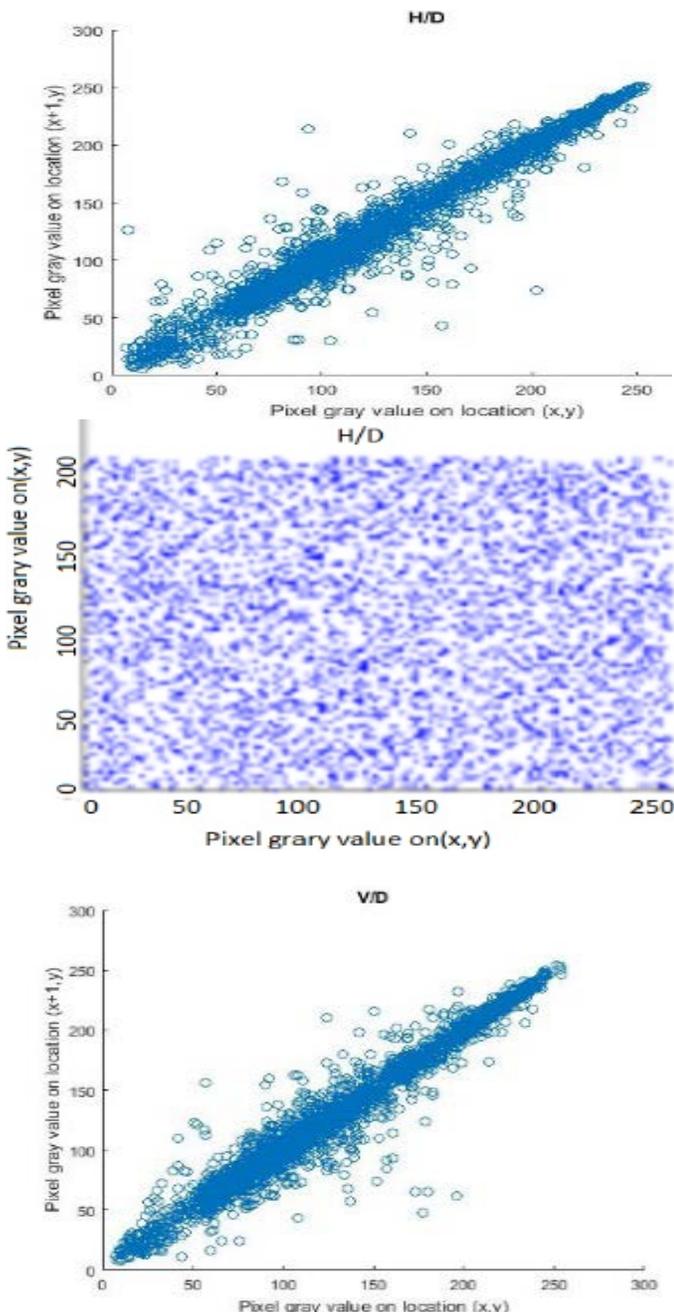


Fig. 14. Correlation Test of the RGB Components of the Plain Image and Ciphered Image. The Figure Displays the Distribution Coefficient of nearby Pixels in Three Directions (H/D, V/D, and D/D) for the Plain Image (up) and the Ciphered Image (down).

#### D. Discovery of Differential Attack

To evaluate the capacity to withstand a differential attack, the Pixel Change Rate Number, and the Unified Average Change Intensity (UACI) tests are applied. The NPCR test measures the number of different pixels between plain and encrypted images, and the UACI test measures the average intensity of these two images [29]. A gentle pixel change generates a softly modified cipher image. Analyzing the relationship between the cipher image created and the plain image using NPCR and UACI becomes essential. The definition of UACI and NPCR is as follows:

$$UACI_{R,G,B} = \left( \frac{\sum_{ij} |c_{R,G,B}(i,j) - c'_{R,G,B}(i,j)|}{255} \right) \times 100 \quad (21)$$

$$NPCR_{R,G,B} = \left( \frac{\sum_{ij} D_{R,G,B}(i,j)}{M \times N} \right) \times 100 \quad (22)$$

$$D_{R,G,B}(i,j) = \begin{cases} 0 & \text{if } C_{R,G,B}(i,j) = C'_{R,G,B}(i,j) \\ 1 & \text{Otherwise} \end{cases} \quad (23)$$

Respectively, the width and height of the image are M and N, and the pixel values in the *i*-th row and *j*-th column are  $C_{R,G,B}(i,j)$  and  $C'_{R,G,B}(i,j)$  for the two ciphered images before and after one pixel of the original plain image is altered. Table IV shows the values of  $NPCR_{R,G,B}$  over 99.55% and values of  $UACI_{R,G,B}$  above 33.44%. The studies illustrate that our technique is highly sensitive to minor changes in the original image, even if the two original images are only one-bit different, the decrypted images are somewhat different.

TABLE IV. RESULTS OF NPCR AND UACI (PERCENT)

| Image  | The Proposed Algorithm |         |         | Ref [10] |         |
|--------|------------------------|---------|---------|----------|---------|
|        |                        | NPCR%   | UACI %  | NPCR%    | UACI %  |
| Lena   | R                      | 99.5117 | 33.4218 | 99.6052  | 33.4132 |
|        | G                      | 99.4751 | 33.4411 |          |         |
|        | B                      | 99.5132 | 33.4887 |          |         |
| Baboon | R                      | 99.5895 | 33.6405 | 99.6227  | 33.4865 |
|        | G                      | 99.5728 | 33.3403 |          |         |
|        | B                      | 99.6368 | 33.4706 |          |         |

#### E. Peak Signal-to-Noise Ratio (PSNR) Analysis

In image reconstruction, PSNR is used primarily as a quality metric. The following equation is calculated:

$$PSNR_{R,G,B} = 20 * \log_{10} \left( \frac{255}{\sqrt{MSE_{R,G,B}}} \right) \quad (24)$$

$$MSE_{R,G,B} = \sum_i \sum_j \frac{C_{R,G,B}(i,j) - C'_{R,G,B}(i,j)}{M \times N} \quad (25)$$

The mean square error (MSE) describes the difference in the values from 0 to 255 between the plain and the ciphered image. The variations between the original and the encrypted image in the PSNR values also are shown in Table V. Our approach shows higher resistance to statistical attacks.

TABLE V. RESULTS OF PSNR

| Image  | New algorithm |        |        | Ref [27] |        |        |
|--------|---------------|--------|--------|----------|--------|--------|
|        | R             | G      | B      | R        | G      | B      |
| Lena   | 7.9687        | 8.8887 | 9.7595 | 7.8992   | 8.5765 | 9.6785 |
| Baboon | 8.3985        | 9.4578 | 8.9701 | 8.9581   | 9.4143 | 8.4156 |

## VIII. CONFLICT OF INTEREST

The authors declare that there is no conflict of interest and there is no competition in the financial interest.

## IX. CONCLUSION

We proposed a large enough key space algorithm to resist brute-force attacks for image security. The proposed cryptosystem generates three keys in the form of a square matrix  $M \times N$ , and three keys in the form of a vector matrix of length  $MN$ . Our cryptosystem for image encryption technique is based on a combination of multidimensional chaos systems and the diversity between shuffling and scrambling of rows and columns in the RGB components of the plain image, as well as multi-level diffusion of pixel values. The computational study between the proposed algorithm and other cryptosystem showed that the proposed algorithm has high level of security and wide key spaces. The advantage of the current algorithm is using multi-level encryption based on big key space.

## REFERENCES

- [1] Mohammed A. B. Younes, "Literature Survey on Different Techniques of Image Encryption", International Journal of Scientific & Engineering Research, Vol. 7, Issue 1, 2016.
- [2] Robert A J Matthews "On the derivation of a chaotic encryption algorithm" J. Cryptologia, Vol. 13, No. 1, 1989, pp. 29-42.
- [3] Chanil Pak, Lilian Huang" A new color image encryption using combination of the ID chaotic map", Signal Processing 138 (2017) 129–137.
- [4] C. Fu, Z. Zhang and Y. Cao, "An Improved Image Encryption Algorithm Based on Chaotic Maps," Third International Conference on Natural Computation (ICNC 2007), Vol. 5, 2007, pp. 24-27.
- [5] LIU Xiangdong, Zhang Junxing, Zhang Jinhai, He Xiqin." Image Scrambling Algorithm Based on Chaos Theory and Sorting Transformation". IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.1, January 2008.
- [6] Juan Li, Yong Feng, Xuqiang Yang" Discrete Chaotic based 3D Image encryption Scheme", Symposium on Photonics and Optoelectronics, September 2009 IEEE.
- [7] Qais H. Alsafasfeh, Aouda A. Arfoa," Image Encryption Based on the General Approach for Multiple Chaotic Systems", Journal of Signal and Information Processing, 2011, 2, 238-244.
- [8] H. Zhang, X.F. Wang, Z.H. Li, and D.H. Liu, "A Fast Image Encryption Algorithm Based on Chaos System and Henon Map", Journal of computer Research and Development, Vol. 42, issue 12, 2137-2142, 2005.
- [9] Seyed Mohammad Seyedzadeh, Sattar Mirzakuchaki "A fast color image encryption algorithm based on coupled two-dimensional piecewise chaotic map" Signal Processing 92 (2012) 1202–1215.
- [10] Guoji Zhang a, Qing Liu b," A novel image encryption method based on total shuffling scheme", Optics Communications 284 (2011) 2775–2780.
- [11] Hongjuan Liu, Zhiliang Zhu, Huiyan Jiang, and Beilei Wang," A Novel Image Encryption Algorithm Based on Improved 3D Chaotic Cat Map", The 9th International Conference for Young Computer Scientists 978-0-7695-3398, 2008 IEEE.
- [12] Howard Anton and Chris Rorres "Elementary linear Algebra, Applications Version", Tenth Edition, WILEY, 2011.
- [13] Ibrahim Yasser, Fahmi Khalifa, Mohamed A. Mohamed, and Ahmed S. Samrah" A New Image Encryption Scheme Based on Hybrid Chaotic Maps" J. Complexity Volume 2020, Article ID 9597619, 23 pages.
- [14] M. Kumari, S. Gupta, P. Sardana, "A survey of image encryption algorithms", 3D Research, Volume 8, Number 4, (2017) Article No.:148
- [15] P. N. Khade and M. Narnaware," 3D Chaotic Functions for Image Encryption", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

- [16] Xingyuan Wang, LinTeng, Xue Qin, "A novel colour image encryption algorithm based on chaos", *Signal Processing* 92 (2012) 1101–1108.
- [17] C. Hoppen, Y. Kohayakawa, C. Moreira, B. Ráth, R. Sampaio "Limits of permutation sequences", *Journal of Combinatorial Theory, S. B* 103 (2013) 93–113.
- [18] W. El-Shafai, et al. "Robust medical image encryption based on DNA-chaos cryptosystem for secure telemedicine and healthcare applications", *Journal of Ambient Intelligence and Humanized Computing*, March 2021.
- [19] S. Mazloom and A.M. Eftekhari-Moghadam, "Colour image encryption based on coupled nonlinear chaotic map", *Chaos, Solitons & Fractals* 42 (3) (2009) 1745–1754.
- [20] Sudhir Keshari, Dr. S. G. Modani, "Image Encryption Algorithm based on Chaotic Map Lattice and Arnold cat map for Secure Transmission", *IJCST Vol. 2, Issue 1, March 2011*.
- [21] J. Wu, X. Liao, and B. Yang, "Image encryption using 2D Hénon-Sine map and DNA approach", *Signal Processing*, 2018, vol. 153, pp.11–23.
- [22] C. Wu, Y. Wang, Y. Chen, J. Wang, and Q. Wang, "Asymmetric encryption of multiple-image based on compressed sensing and phase-truncation in cylindrical diffraction domain," *Optics Communications*, 2019, vol. 431, 203–209.
- [23] C.-Y. Song, Y.-L. Qiao, and X.-Z. Zhang, "An image encryption scheme based on new spatiotemporal chaos," *Optik-International Journal for Light and Electron Optics*, 2013, vol. 124, no. 18, pp. 3329–3334.
- [24] C. c. chang, M. Hwang, T. chen." A new encryption algorithm for image cryptosystem", *Journal of system and software*, vol. 58, 2001, 83-91.
- [25] X. Wang and L. Teng, "A bit-level image encryption algorithm based on spatiotemporal chaotic system and self-adaptive", *Optics Communications Volume 285, Issue 20, 15 September 2012, Pages 4048-4054*.
- [26] Zhenjun Tang, Ye Yang, Shijie Xu, Chunqiang Yu, and Xianquan Zhang" Image Encryption with Double Spiral Scans and Chaotic Maps" *Security and Communication Networks Volume 2019, Article ID 8694678, 15 pages*.
- [27] NF Elabady, MI Moussa, HM Abdalkader, SF Sabbeh" Image Encryption Based on New One-Dimensional Chaotic Map", *International Conference on Engineering and Technology (ICET)*, 19-20 April 2014, Cairo, Egypt.
- [28] Arslan Shafique, Mohammad Mazyad Hazzazi, Adel R. Alharbi, Iqtadar Hussain, "Integration of Spatial and Frequency Domain Encryption for Digital Images", *Access IEEE*, vol. 9, pp. 149943-149954, 2021.
- [29] Yue Wu, Joseph P. Noonan, and Sos Agaian, "NPCR and UACI randomness tests for image encryption". *Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT)*, April Edition, 2011.

#### AUTHORS' PROFILE



**Mahmoud I. Moussa** is an associate professor of computer science at the faculty of computers & artificial intelligence, Benha University in Egypt. He received his Ph.D. in Parallel Algorithms (mainly in parallel graph algorithms) from faculty of informatics at Karlsruhe Institute of Technology-KIT, Germany. His research interests span both theoretical computer science and information security. Much of his work has been on improving the understanding, design, and performance of algorithms and analysis, mainly through the application of graph algorithms, bioinformatics, as well as Steganography and Cryptography. Moussa's work includes a prediction method for biological activity using random forests and kernel functions in Support Vector Machine (SVM), image encryption using chaotic maps, a method for using smartphone devices efficiently and offloading only when necessary.



**Eman I. Abd El-Latif** received the M.Sc. degree and Ph.D. in computer science, at Faculty of Science, Benha University, Egypt, 2016 and 2020, respectively. She is currently working as lecturer at mathematics department, Benha University, Egypt. Her areas of research include Digital Forensics, Security (Encryption – Steganography), and image processing.



**Nawaz Majid** is an assistant professor of computer science at Department computer science, Faculty of science, Northern Border University (NBU), KSA. His research interests span both theoretical computer science and information security. Adding more to that, he worked on research including Data mining and knowledge discovery.

# A Visual-Range Cloud Cover Image Dataset for Deep Learning Models

Muhammad Umair<sup>ID</sup>, Manzoor Ahmed Hashmani<sup>ID</sup>

High Performance Cloud Computing Center (HPC3)  
Department of Computer and Information Sciences  
Universiti Teknologi PETRONAS  
Seri Iskandar, Malaysia

**Abstract**—Coastal and offshore oil and gas structures and operations are subject to continuous exposure to environmental conditions (ECs) such as varying air and water temperatures, rough sea conditions, strong winds, high humidity, rain, and varying cloud cover. To monitor ECs, weather and wave sensors are installed on these facilities. However, the capital expenditure (CAPEX) and operational expenses (OPEX) of these sensors are high, especially for offshore structures. For observable ECs, such as cloud cover, a cost-effective deep learning (DL) classification model can be employed as an alternative solution. However, to train and test a DL model, a cloud cover image dataset is required. In this paper, we present a novel visual-range cloud cover image dataset for cloud cover classification using a deep learning model. Various visual-range sky images are captured on nine different occasions, covering six cloud cover conditions. For each cloud cover condition, 100 images are manually classified. To increase the size and quality of images, multiple label-preserving data augmentation techniques are applied. As a result, the dataset is expanded to 9,600 images. Moreover, to evaluate the usefulness of the proposed dataset, three DL classification models, i.e., GoogLeNet, ResNet-50, and EfficientNet-B0, are trained, tested, and their results are presented. Even though EfficientNet-B0 had better generalization ability and marginally higher classification accuracy, it was discovered that ResNet-50 is the best choice for cloud cover classification due to its lower computational cost and competitive classification accuracy. Based on these results, it is concluded that the proposed dataset can be used in further research in DL-based cloud cover classification model development.

**Keywords**—Cloud cover; dataset; classification; GoogLeNet; ResNet-50; EfficientNet-B0

## I. INTRODUCTION

Cloud cover is an important observation for weather monitoring. It is classified as a percentage cloud cover of the visible sky. Changes in cloud cover percentage affect the global mean surface temperature and pressure systems [1], the amount of solar UV radiation reaching the earth's surface [2], the melting rate of ice shields [3], and the radiation-energy-carbon balance of tropical rain forests [4]. It also plays a crucial part in the selection of sites for observational astronomy [5, 6]. From the clean power generation perspective, cloud cover percentage has an instantaneous effect on the power generation capability of solar panels [7]. For safety reasons, offshore oil and gas excavation and production activities are

also subject to weather conditions [8], which may be associated with cloud cover.

Since cloud cover is an important parameter of prevailing weather conditions, its classification plays an important role in weather monitoring at offshore oil and gas platforms. Currently, these platforms rely on various sensors to monitor weather conditions. The procurement cost of these sensors is usually high. Additionally, any subsequent maintenance activity is also costly due to the remoteness of the offshore site. As a low-cost alternative, deep learning (DL)-based technologies can be applied to monitor, forecast, and predict climate and weather conditions [9], and postprocess cloud cover [10]. A DL-based cloud cover monitoring system requires the installation of a less expensive visual-range sensor, a single-board computer, and a pre-trained DL classification model. This solution can be applied as support to the existing sensor-based system or deployed as a module for a larger DL-based weather monitoring system at remote oil and gas platforms.

The deep learning classification models depend upon a large collection of images for training and testing purposes. Publicly available weather image datasets do not distinguish between different cloud cover conditions [11-13]. As a result, they are not suitable for a multi-class cloud cover classification problems. The objective of this study is to fill this gap by proposing a novel visual-range cloud cover image dataset, named Manzoor-Umair: Cloud Cover Dataset (MU-CCD), for deep learning classification models. The proposed study classifies cloud cover conditions into six classes. For every cloud cover condition, 100 source images are manually identified. Various augmentation techniques are applied to the source images to improve the quality and quantity of each cloud cover condition. As a result, the dataset consists of 9,600 RGB images at 1920x1080 pixel dimensions.

The dataset is aimed at a DL-based classification module of cloud cover for a larger DL-based weather classification system to be deployed at remote oil and gas facility. Thus, to access its suitability, the MU-CCD is evaluated on three well-known deep learning image classification models, namely GoogLeNet, ResNet-50, and EfficientNet-B0. The presented results indicate that the proposed dataset is well suited for training and testing of deep learning-based cloud cover classification models and, as a result, can be used for the

This research work is a part of an ongoing research project funded by Yayasan Universiti Teknologi PETRONAS – Fundamental Research Grant (YUTP-FRG-2019, grant number 015LC0-158).

development of cloud cover classification modules of a larger DL-based weather monitoring system.

The rest of the paper is divided in the following manner: The literature review section describes the related literature on cloud cover classification and publicly available weather image datasets. The methodology section explains the steps taken to create MU-CCD. The proposed dataset section discusses the different features and statistics of MU-CCD. The dataset effectiveness experiment section presents the classification performance of well-known DL-based image classification models on MU-CCD, and finally, the conclusion and future work section sums up the work and identifies the future directions.

## II. LITERATURE REVIEW

### A. Cloud Cover Classification

The most common way to classify cloud cover is by visual observation of the sky by an experienced meteorologist [14]. This method divides the sky into eight segments, called octas (C). For every segment, the percentage presence of cloud cover is noted. The cumulative percentage cover of clouds then identifies the present cloud cover conditions. The method has been used in studies conducted by Robaa [15] and Werkmeister et al. [16]. Other methods of identifying cloud cover include the processing of images from all-sky cameras and satellite data. However, based on its simplicity and practicability, classifying cloud cover conditions by observing sky images for cloud cover percentage is found to be more suitable and relevant for the presented work.

### B. Visual-Range Weather Image Datasets

In this section, recently published image weather datasets are selected to assess their suitability for cloud cover classification problems. Since the presented study focuses on visual-range cloud cover image classification, thus the selected datasets consist of visual-range weather images under various geographical and weather conditions.

The RFS Weather Dataset is a visual-range image dataset aimed at computer vision applications [11]. The dataset is a collection of images acquired from various online resources such as Creative Commons, Flickr, Pixabay, and Wikimedia Commons. The images in the data are divided into three categories, namely, rain, fog, and snow. Additionally, the dataset borrowed images of sunny and cloudy categories from a dataset proposed by Lu et al. [17]. For each category, there are 1,100 images. Thus, in total, the RFS Weather Dataset consists of 5,500 images. However, the cloudy images in the dataset are not categorized by the percentage of cloud cover. Thus, to classify different cloud cover conditions, this dataset is found to be unsuitable for training and testing of a deep learning classification model.

The 4Seasons is a multi-weather visual-range video dataset aimed at autonomous vehicle driving applications [12]. The dataset covers three weather conditions, namely, sunny, overcast, and snowy, as well as two illumination conditions, namely, day and night. Due to its application nature, the dataset does not further bifurcate the overcast conditions into various cloud cover classes. The absence of such bifurcation makes it

unsuitable for cloud cover classification using deep learning models.

The Image2Weather dataset is a large-scale visual-range image dataset aimed at weather conditions and temperature estimations [13]. The dataset consists of 180,000 images and covers five weather types, namely, sunny, cloudy, foggy, rainy, and snowy. The dataset was created using existing images from an online resource. Based on the image metadata, its geographical location and image capture time were identified. This information is then used to retrieve corresponding weather information from an online weather center. However, the cloudy weather in the presented dataset is not further categorized into different cloud cover conditions, which makes it unsuitable for training and testing of deep learning classification models.

Based on the presented evidence, it is deduced that, at present stage, the reviewed weather image datasets are unsuitable for machine classification of cloud cover conditions. For example, all three datasets, i.e., the RFS Weather Dataset [11], 4Seasons dataset [12], and Image2Weather dataset [13], do not categorize cloud cover images as a percentage of sky cover. Thus, classification of various cloud cover conditions is not possible using these datasets. Thus, in this paper, to address the need for training and testing dataset for deep learning cloud cover classification models, we propose a visual-range cloud cover image dataset that presents six cloud cover conditions based on percentage of sky cover.

## III. METHODOLOGY

This section discusses the optical sensor and methodology of MU-CCD development.

### A. Optical Sensor

In this experiment, a visual-range 24 mega-pixel NIKON D3400 camera is utilized to capture images in JPEG format at 6000 x 4000 pixels in sRGB color space. The camera was set to auto mode and various zoom levels were applied throughout the data collection process.

### B. Data Collection and Preprocessing

In the first phase of Manzoor-Umair: Cloud Cover Dataset (MU-CCD) development, sky images were captured on nine different occasions across five geographically separated locations in West Malaysia. The images are taken across the year to capture different seasonal attributes. In addition to this, to record varying levels of illumination, the images are taken at different times of the day. These images are visually analyzed and sub-images containing sky conditions are extracted at a resolution of 1920x1080 pixels. Images having undesired artefacts or sensor noise are then removed. The highest visual quality images are selected for the second phase.

### C. Image Classification

The second phase of MU-CCD creation is the manual classification of selected images. We have classified cloud cover conditions into six categories. These classes are clear sky, few clouds, isolated clouds, scattered clouds, broken clouds, and overcast. Table I describes the empirical methodology adopted for CC classification based on cloud

cover percentage in an image. For each class, 100 images are manually classified.

#### D. Data Augmentation

Data augmentation is an effective technique to increase the quality and size of image datasets. It not only provides a solution for limited data but also addresses the issue of overfitting in DL models [18]. In the third phase, a data augmentation policy (DAP) is designed to augment the image data. The policy makes sure that label-preserving augmentation techniques such as flipping, down sampling, color space transformation, noise injection, grid shuffle, weather-related augmentation, and kernel filtering are applied to images. Based on the DAP, an image augmentation pipeline (IAP) is developed using a Python-based library, Albumentations [19],

and applied to all six source image pools. This results in 1,600 images per class. The flow of IAP is illustrated in Fig. 1.

TABLE I. CLOUD COVER CLASSIFICATION

| Class ID | Class Name       | Visible Cloud Cover (%) |
|----------|------------------|-------------------------|
| 1        | Clear Sky        | 0                       |
| 2        | Few Clouds       | 1-10                    |
| 3        | Isolated Clouds  | 11-25                   |
| 4        | Scattered Clouds | 26-50                   |
| 5        | Broken Clouds    | 51-90                   |
| 6        | Overcast         | 91-100                  |

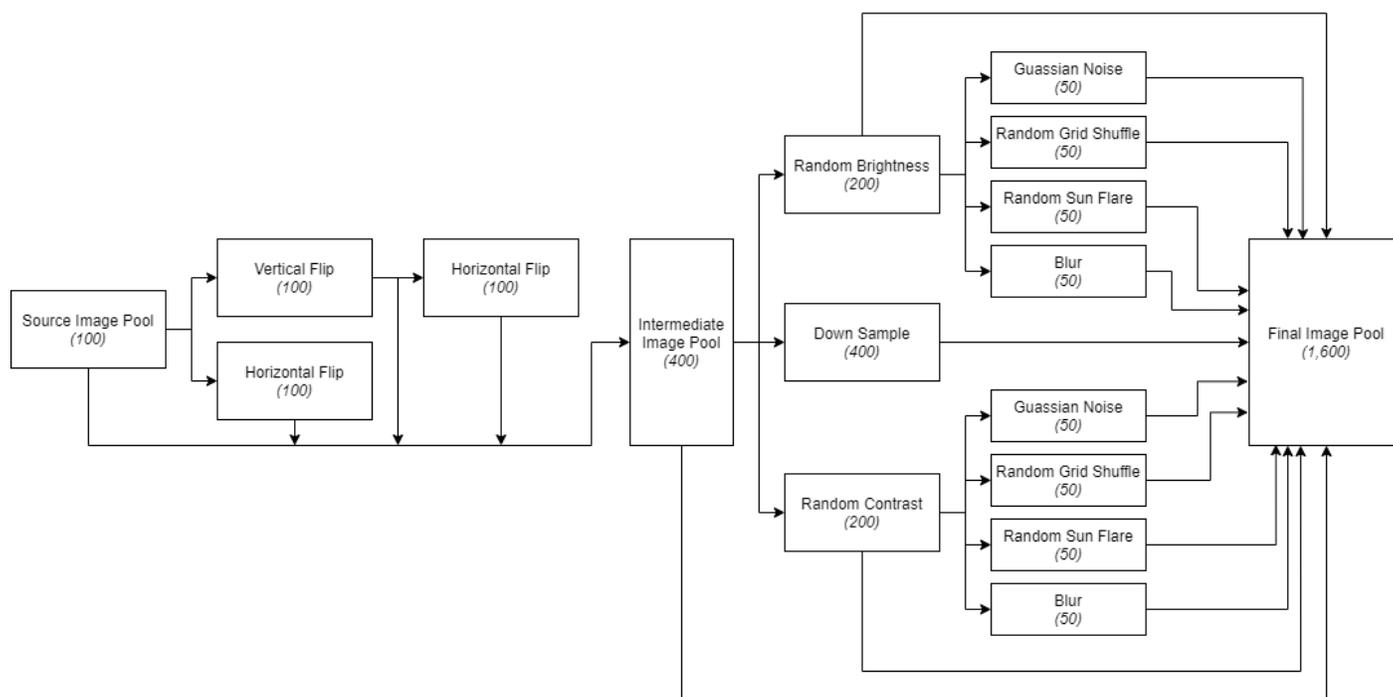


Fig. 1. Image Augmentation Pipeline (IAP).

### E. Image Naming Convention

After each stage of IAP, a suffix is added to the output image's file name. The list of used suffixes and their corresponding augmentation methods is given in Table II.

TABLE II. AUGMENTATION SETS AND CORRESPONDING FILE NAME SUFFIXES

| Serial No. | Augmentation Set | Augmentation Method | Suffix |
|------------|------------------|---------------------|--------|
| 1          | AS1              | Horizontal Flip     | HF     |
| 2          |                  | Vertical Flip       | VF     |
| 3          | AS2              | Down Sample         | DS     |
| 4          | AS3              | Random Brightness   | RB     |
| 5          |                  | Random Contrast     | RC     |
| 6          | AS4              | Gaussian Noise      | GN     |
| 7          |                  | Random Grid Shuffle | GS     |
| 8          |                  | Solar Flare         | SL     |
| 9          |                  | Blur                | BL     |

A sample file called "DSC\_0183\_7\_VF\_RC\_BL.JPG" indicates that the source image is "DSC\_0183\_7.JPG" and that it has been vertically flipped (VF), randomly contrasted (RC), and blurred (BL).

### IV. PROPOSED DATASET

MU-CCD is a visual-range image dataset of six cloud cover classes. The dataset is designed for training and testing of DL-

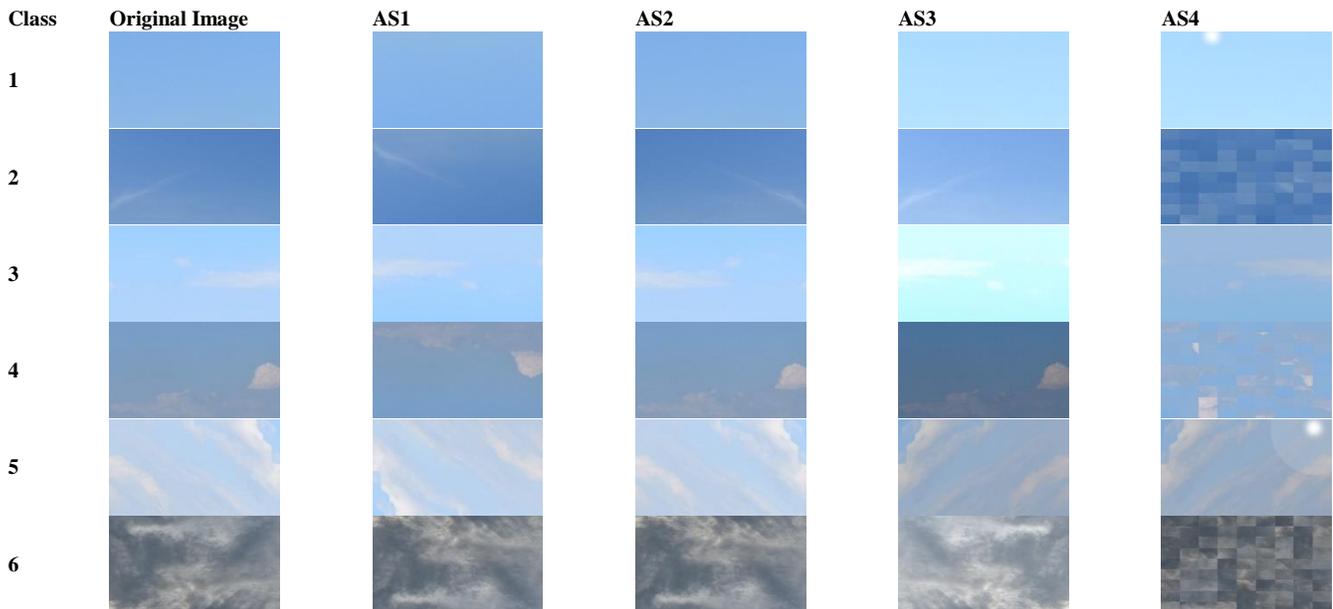


Fig. 2. MU-CCD Class-Wise Original and Augmented Image Samples.

based cloud cover classification problems. The presented image format is JPG in RGB color space, and the dimensions are 1920x1080 pixels. By applying nine different augmentation methods, the number of image instances per class is increased from 100 to 1,600, resulting in a total of 9,600 images in the dataset. The images in final dataset are then randomly divided into training and testing sets at a ratio of 80:20. The summary of MU-CCD is presented in Table III.

TABLE III. SUMMARY OF MU-CCD

| Source Images per Class | Augmentation Methods Applied | Final Images per Class | Total Images in Dataset | Training Images | Testing Images |
|-------------------------|------------------------------|------------------------|-------------------------|-----------------|----------------|
| 100                     | 9                            | 1,600                  | 9,600                   | 7,680           | 1,920          |

The dataset takes advantage of various data augmentation techniques to increase its dataspace. As a result of the carefully designed DAP, it represents a balanced proportion of image augmentation sets. For example, the dataset can be divided into four image augmentation sets. A tabular description is presented in Table II. The first augmentation set (AS1) consists of the original image and its flipped versions. The second set (AS2) contains down-scaled images. The third augmentation set (AS3) pool has images with color space variations. Finally, the fourth augmentation set (AS4) has a mix of noise injection, grid shuffle, weather-related augmentation, and kernel filtering applied to the images. Each set has an equal proportion of 25% in the dataset.

For the interest of the reader, in Fig. 2, class-wise original and augmented image samples from MU-CCD are presented.

## V. DATASET EFFECTIVENESS EXPERIMENT

To validate the effectiveness of MU-CCD in DL-based cloud cover classification problems and to provide reference measures, we have trained and tested three selected DL models on the proposed dataset. In the following subsections, the hardware and software setup, selection of deep learning models, experimental configurations, training and validation results, and discussion are presented.

### A. Hardware and Software Setup

The experiment was conducted on an Intel (R) Core (TM) i7-9750H CPU running at 2.60GHz. The machine has a 16 GB main memory and runs the Windows 10 Pro operating system. An NVIDIA GeForce RTX 2070 with a Max-Q Design GPU is used for DL model training and testing. The GPU has 8 GB of memory. All the simulations are conducted on MATLAB version R2021a.

### B. Deep Learning Classification Models' Selection

For the proposed experiments, three DL models are selected based on their ability to improve computational accuracy, ease of training, and effective convolutional neural networks (CNNs) network scaling ability. The selected models are GoogLeNet [20], ResNet-50 [21], and EfficientNet-B0 [22].

The GoogLeNet network was developed by Google [20]. It is the winner of the ILSVRC 2014 competition. The network is based on the Inception architecture, and its receptive field size is 224x224 pixels in RGB color space. GoogLeNet is a 22-layer deep network that focuses on improving computational accuracy.

Deep residual nets, or ResNet, was the winner of the ILSVRC 2014 competition [21]. The plain network architecture is inspired by the VGG nets. There are different variants of ResNet based on the depth of its layers. The 50-layer deep variant is known as ResNet-50. The network has a receptive field size of 224x224 pixels in RGB color space. The network is based on a residual learning framework that eases the training of deeper neural networks.

The EfficientNet-B0 model is developed by Google. The baseline network is built using neural architecture search, which optimizes the accuracy and efficiency of the network [22]. The EfficientNet-B0 has 290 layers, and its receptive field size is 224x224 pixels in RGB color space. The model has various variants and has shown improvement in the top-1 accuracy for ResNet-50 on the ImageNet dataset.

### C. Experimental Configurations

Three separate cloud cover classification experiments are designed and performed on MU-CCD. For all three experiments, Stochastic Gradient Descent with momentum (SGDM) was selected as an optimization algorithm as it is known for its faster convergence. The initial learning rate for SGDM is set at 0.01. The validation frequency and maximum epoch number are fixed at 50 and 10, respectively. Due to the relatively increased layer depth of the ResNet-50 and EfficientNet-B0 models and GPU memory constraints, an image batch size of 64 is selected for these models. The GoogLeNet, however, is trained using an image batch size of

128. For all three experiments, the validation patience is set at 5.

### D. Training, Validation Results and Discussion

All three models are trained and validated on a training-validation set (TVS) of MU-CCD. The TVS has 7,680 images, which are split into 70% and 30% for training and validation purposes. The training accuracy and loss graphs for all three models are presented in Fig. 4, 5, and 6. In each figure, the top graph depicts training accuracy performance, while the bottom graph represents the corresponding loss. The legends for Fig. 4, 5, and 6 are presented in Fig. 3.

The GoogLeNet took the least amount of time (i.e., 415 seconds) to train the network, and attained the highest training accuracy of 98.4%. However, its validation accuracy remains the lowest in the group. Fig. 4 illustrates the training accuracy and loss for GoogLeNet.

As depicted in Fig. 5 and 6, the ResNet-50 and EfficientNet-B0 models' validation accuracies were almost similar and slightly higher than GoogLeNet. However, when compared for training time, ResNet-50 was trained 27% faster than EfficientNet-B0.

For all three models, overfitting is observed, as their validation loss remains higher than their training loss. Moreover, it was observed that the EfficientNet-B0 generalized well, as it yielded the lowest difference between its training and validation loss. Table IV presents the training and validation statistics of the evaluated models.

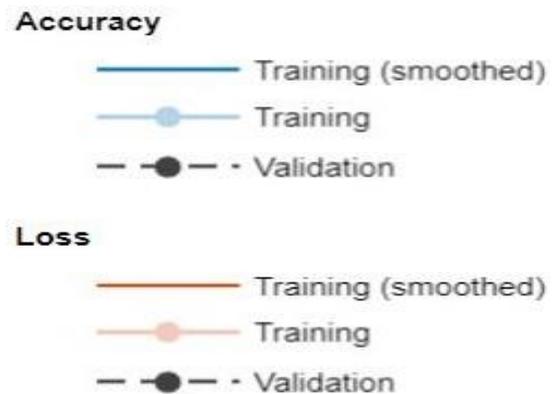


Fig. 3. Legdens for Fig. 4, 5 and 6.

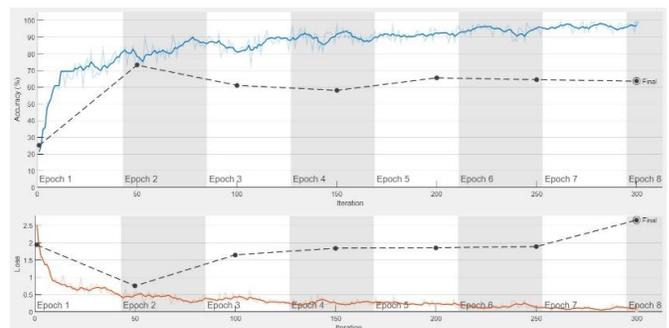


Fig. 4. Training (Top), Validation (Bottom) Graphs for GoogLeNet.

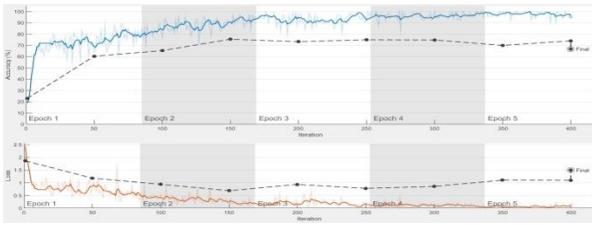


Fig. 5. Training (Top), Validation (Bottom) Graphs for ResNet-50.

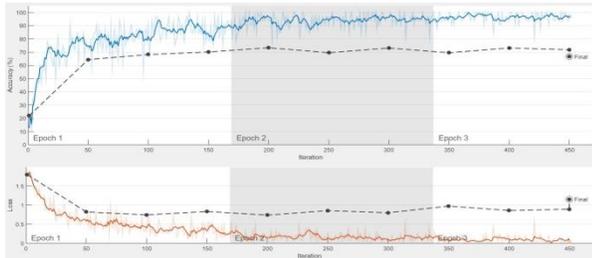


Fig. 6. Training (Top), Validation (Bottom) Graphs for EfficientNet-B0.

**E. Testing Results and Discussion**

All three models are tested on the testing set (TS) of MU-CCD, which consists of 1,920 images.

During the training-validation phase, it was found that EfficientNet-B0 generalized well as compared to ResNet-50 and GoogLeNet. By analyzing the confusion matrix for GoogLeNet, ResNet-50, and EfficientNet-B0 in Fig. 7, 8, and 9, respectively, a similar observation is made. EfficientNet-

B0’s overall classification accuracy of 87.2% remains the highest among all three models. However, it was marginally higher than ResNet-50’s classification accuracy. For 3 classes (few clouds, isolated clouds, and overcast), EfficientNet-B0 resulted in the highest classification accuracy. The scaled-up network dimensions of EfficientNet-B0 can be attributed to a higher processing time per image during the testing phase. It was 3 times higher than the next best model, i.e., ResNet-50. The confusion matrix for EfficientNet-B0 is presented in Fig. 7.

The ResNet-50 showed very competitive results among all the deep learning models in question. The model not only took less time to process an image, but it also classified the cloud cover with high accuracy. Despite the poor performance of GoogLeNet and EfficientNet-B0 for class 4 (Scattered Clouds) instances’ classification, ResNet-50 yielded relatively better results. As presented in Table V, across all six classes, the model’s classification performance remains highly competitive with the other two models. The confusion matrix for ResNet-50 is presented in Fig. 8 which shows an overall accuracy of 87.0%.

Due to its simple architecture, the GoogLeNet took the least amount of time to process an image during the testing phase. However, this performance is marginally better than ResNet-50. Similarly, in classifying class 1 and 6 instances, the model slightly surpassed ResNet-50, but the model’s over-all classification performance remains the lowest among all models. The model’s overall accuracy was 83.5%, and it is illustrated in Fig. 9.

TABLE IV. TRAINING AND VALIDATION RESULTS OF DEEP LEARNING MODELS

| Model           | Training Time (sec) | Training Accuracy | Validation Accuracy | Training Loss | Validation Loss |
|-----------------|---------------------|-------------------|---------------------|---------------|-----------------|
| GoogLeNet       | 415                 | 98.43%            | 63.63%              | 0.06          | 2.65            |
| ResNet-50       | 655                 | 95.31%            | 66.75%              | 0.10          | 1.09            |
| EfficientNet-B0 | 895                 | 96.87%            | 66.62%              | 0.04          | 0.89            |

TABLE V. CLASSIFICATION RESULTS OF DEEP LEARNING MODELS

| Model           | Classification Accuracy |         |         |         |         |         |         | Per Image Processing Time (sec) |
|-----------------|-------------------------|---------|---------|---------|---------|---------|---------|---------------------------------|
|                 | Class 1                 | Class 2 | Class 3 | Class 4 | Class 5 | Class 6 | Overall |                                 |
| GoogLeNet       | 100.0%                  | 85.0%   | 65.0%   | 66.9%   | 85.6%   | 98.4%   | 83.5%   | 0.09                            |
| ResNet-50       | 97.5%                   | 85.9%   | 77.2%   | 73.4%   | 90.9%   | 97.2%   | 87.0%   | 0.11                            |
| EfficientNet-B0 | 98.4%                   | 89.7%   | 88.1%   | 69.4%   | 77.5%   | 100.0%  | 87.2%   | 0.32                            |

| Output Class | 1             | 2              | 3              | 4              | 5              | 6            | Accuracy       |
|--------------|---------------|----------------|----------------|----------------|----------------|--------------|----------------|
| 1            | 315<br>16.4%  | 13<br>0.7%     | 1<br>0.1%      | 0<br>0.0%      | 0<br>0.0%      | 0<br>0.0%    | 95.7%<br>4.3%  |
| 2            | 5<br>0.3%     | 287<br>14.9%   | 34<br>1.8%     | 4<br>0.2%      | 0<br>0.0%      | 0<br>0.0%    | 87.0%<br>13.0% |
| 3            | 0<br>0.0%     | 18<br>0.9%     | 282<br>14.7%   | 90<br>4.7%     | 14<br>0.7%     | 0<br>0.0%    | 69.8%<br>30.2% |
| 4            | 0<br>0.0%     | 2<br>0.1%      | 3<br>0.2%      | 222<br>11.6%   | 24<br>1.2%     | 0<br>0.0%    | 88.4%<br>11.6% |
| 5            | 0<br>0.0%     | 0<br>0.0%      | 0<br>0.0%      | 3<br>0.2%      | 248<br>12.9%   | 0<br>0.0%    | 98.8%<br>1.2%  |
| 6            | 0<br>0.0%     | 0<br>0.0%      | 0<br>0.0%      | 1<br>0.1%      | 34<br>1.8%     | 320<br>16.7% | 90.1%<br>9.9%  |
|              | 98.4%<br>1.6% | 89.7%<br>10.3% | 88.1%<br>11.9% | 69.4%<br>30.6% | 77.5%<br>22.5% | 100%<br>0.0% | 87.2%<br>12.8% |

Fig. 7. EfficientNet-B0 Confusion Matrix.

| Output Class | 1             | 2              | 3              | 4              | 5             | 6             | Accuracy       |
|--------------|---------------|----------------|----------------|----------------|---------------|---------------|----------------|
| 1            | 312<br>16.2%  | 34<br>1.8%     | 2<br>0.1%      | 0<br>0.0%      | 0<br>0.0%     | 0<br>0.0%     | 89.7%<br>10.3% |
| 2            | 8<br>0.4%     | 275<br>14.3%   | 63<br>3.3%     | 16<br>0.8%     | 0<br>0.0%     | 0<br>0.0%     | 76.0%<br>24.0% |
| 3            | 0<br>0.0%     | 10<br>0.5%     | 247<br>12.9%   | 65<br>3.4%     | 8<br>0.4%     | 5<br>0.3%     | 73.7%<br>26.3% |
| 4            | 0<br>0.0%     | 1<br>0.1%      | 8<br>0.4%      | 235<br>12.2%   | 10<br>0.5%    | 3<br>0.2%     | 91.4%<br>8.6%  |
| 5            | 0<br>0.0%     | 0<br>0.0%      | 0<br>0.0%      | 2<br>0.1%      | 291<br>15.2%  | 1<br>0.1%     | 99.0%<br>1.0%  |
| 6            | 0<br>0.0%     | 0<br>0.0%      | 0<br>0.0%      | 2<br>0.1%      | 11<br>0.6%    | 311<br>16.2%  | 96.0%<br>4.0%  |
|              | 97.5%<br>2.5% | 85.9%<br>14.1% | 77.2%<br>22.8% | 73.4%<br>26.6% | 90.9%<br>9.1% | 97.2%<br>2.8% | 87.0%<br>13.0% |

Fig. 8. ResNet-50 Confusion Matrix.

| Output Class | 1            | 2              | 3              | 4              | 5              | 6             | Accuracy       |
|--------------|--------------|----------------|----------------|----------------|----------------|---------------|----------------|
| 1            | 320<br>16.7% | 32<br>1.7%     | 2<br>0.1%      | 0<br>0.0%      | 0<br>0.0%      | 0<br>0.0%     | 90.4%<br>9.6%  |
| 2            | 0<br>0.0%    | 272<br>14.2%   | 102<br>5.3%    | 24<br>1.2%     | 5<br>0.3%      | 0<br>0.0%     | 67.5%<br>32.5% |
| 3            | 0<br>0.0%    | 15<br>0.8%     | 208<br>10.8%   | 74<br>3.9%     | 9<br>0.5%      | 2<br>0.1%     | 67.5%<br>32.5% |
| 4            | 0<br>0.0%    | 1<br>0.1%      | 8<br>0.4%      | 214<br>11.1%   | 14<br>0.7%     | 0<br>0.0%     | 90.3%<br>9.7%  |
| 5            | 0<br>0.0%    | 0<br>0.0%      | 0<br>0.0%      | 5<br>0.3%      | 274<br>14.3%   | 3<br>0.2%     | 97.2%<br>2.8%  |
| 6            | 0<br>0.0%    | 0<br>0.0%      | 0<br>0.0%      | 3<br>0.2%      | 18<br>0.9%     | 315<br>16.4%  | 93.8%<br>6.2%  |
|              | 100%<br>0.0% | 85.0%<br>15.0% | 65.0%<br>35.0% | 66.9%<br>33.1% | 85.6%<br>14.4% | 98.4%<br>1.6% | 83.5%<br>16.5% |

Fig. 9. GoogLeNet Confusion Matrix.

## VI. CONCLUSION AND FUTURE WORK

Installation and maintenance of weather sensors at remote oil and gas platforms entails high CAPEX and OPEX. As an alternative, a low-cost deep learning-based weather monitoring system can be used. One of the components of such a system can be a cloud cover classification model. To train and test this model, in this paper, we have proposed a novel visual-range cloud cover image dataset named MU-CCD. Across West Malaysia, at various occasions and time of the day, sky images were captured and preprocessed. The images were then manually classified into six cloud cover classes. Various label-preserving augmentation techniques were applied on manually classified images to increase the size and quality of the dataset.

As a result, the final dataset consists of 9,600 images covering six cloud cover states. The dataset can be downloaded from <https://www.kaggle.com/umairatwork/manzoorumair-cloud-cover-dataset-muccd>.

The suitability of the proposed dataset for training and testing of the DL classification model was evaluated on three selected DL models. The classification results of these models were also presented. It was found that the dataset is well suited for DL-based cloud cover classification model training and testing. However, it was observed that for clear sky and overcast conditions, the dataset can be further improved for more visually distinct features in the source images. Additionally, it was observed that EfficientNet-b0 generalized well on the presented dataset and effectively classified the images. However, because of its increased number of layers, the model took the longest time to process an image. While considering the classification accuracy and computational cost factors in combination, ResNet-50 emerges as an ideal candidate for the cloud cover classification problem. However, improvement in its generalization capability and classification accuracy can be further investigated as a future work.

## REFERENCES

- [1] M. S. Croke, R. D. Cess, and S. Hameed, "Regional cloud cover change associated with global climate change: Case studies for three regions of the United States," *Journal of Climate*, vol. 12, no. 7, pp. 2128-2134, 1999.
- [2] R. H. Grant and G. M. Heisler, "Effect of cloud cover on UVB exposure under tree canopies: Will climate change affect UVB exposure?," *Photochemistry and photobiology*, vol. 82, no. 2, pp. 487-494, 2006.
- [3] S. Hofer, A. J. Tedstone, X. Fettweis, and J. L. Bamber, "Decreasing cloud cover drives the recent mass loss on the Greenland Ice Sheet," *Science Advances*, vol. 3, no. 6, p. e1700584, 2017, doi:10.1126/sciadv.1700584.
- [4] A. Verhoef, M. S. Moura, and R. Nóbrega, "The effect of cloud cover on the radiation-, energy- and carbon balance of a seasonally dry tropical forest in Brazil (Caatinga)," in *EGU General Assembly Conference Abstracts*, 2020, p. 9210.
- [5] N. Aksaker et al., "Global Site Selection for Astronomy," *Monthly Notices of the Royal Astronomical Society*, vol. 493, no. 1, pp. 1204-1216, 2020, doi: 10.1093/mnras/staa201.
- [6] X. Li, B. Wang, B. Qiu, and C. Wu, "An all-sky camera images classification method using cloud cover features," *Atmospheric Measurement Techniques Discussions*, pp. 1-17, 2021.
- [7] A. Bonkany, S. Madougou, and R. Adamou, "Impacts of Cloud Cover and Dust on the Performance of Photovoltaic Module in Niamey," *Journal of Renewable Energy*, vol. 2017, p. 9107502, 2017/09/07 2017, doi: 10.1155/2017/9107502.
- [8] R. A Halim, M. H. Mhd Yusof, M. H. M Khalid, H. X. Wong, and M. Z. Sulaiman, "Wait on Weather WOW Impact Trending in Malaysia Water: Comprehensive Data Analytics Led to Safe and Optimum Well Planning and Offshore Execution," in *International Petroleum Technology Conference*, 2021, D012S045R138: OnePetro, doi: 10.2523/iptc-21259-ms.
- [9] S. Dewitte, J. P. Cornelis, R. Müller, and A. Munteanu, "Artificial intelligence revolutionises weather forecast, climate monitoring and decadal prediction," *Remote Sensing*, vol. 13, no. 16, p. 3209, 2021.
- [10] F. Dupuy et al., "ARPEGE Cloud Cover Forecast Postprocessing with Convolutional Neural Network," *Weather and Forecasting*, vol. 36, no. 2, pp. 567-586, 2021.
- [11] J. C. V. Guerra, Z. Khanam, S. Ehsan, R. Stolkin, and K. McDonald-Maier, "Weather Classification: A new multi-class dataset, data augmentation approach and comprehensive evaluations of Convolutional Neural Networks," in *2018 NASA/ESA Conference on Adaptive Hardware and Systems (AHS)*, 2018: IEEE, pp. 305-310.

- [12] P. Wenzel et al., "4Seasons: A cross-season dataset for multi-weather SLAM in autonomous driving," in DAGM German Conference on Pattern Recognition, 2020: Springer, pp. 404-417.
- [13] W.-T. Chu, X.-Y. Zheng, and D.-S. Ding, "Image2weather: A large-scale image dataset for weather property estimation," in 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), 2016: IEEE, pp. 137-144.
- [14] E. W. Luiz, F. R. Martins, R. S. Costa, and E. B. Pereira, "Comparison of methodologies for cloud cover estimation in Brazil - A case study," Energy for Sustainable Development, vol. 43, pp. 15-22, 2018, doi: 10.1016/j.esd.2017.12.001.
- [15] S. Robaa, "Evaluation of sunshine duration from cloud data in Egypt," Energy, vol. 33, no. 5, pp. 785-795, 2008, doi: 10.1016/j.energy.2007.12.001.
- [16] A. Werkmeister, M. Lockhoff, M. Schrempf, K. Tohsing, B. Liley, and G. Seckmeyer, "Comparing satellite- to ground-based automated and manual cloud coverage observations – a case study," Atmospheric Measurement Techniques, vol. 8, no. 5, pp. 2001-2015, 2015, doi: 10.5194/amt-8-2001-2015.
- [17] C. Lu, D. Lin, J. Jia, and C. K. Tang, "Two-Class Weather Classification," IEEE Trans Pattern Anal Mach Intell, vol. 39, no. 12, pp. 2510-2524, Dec 2017, doi: 10.1109/TPAMI.2016.2640295.
- [18] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," Journal of Big Data, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0197-0.
- [19] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," Information, vol. 11, no. 2, 2020, doi: 10.3390/info11020125.
- [20] C. Szegedy et al., "Going deeper with convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1-9.
- [21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [22] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, 2019: PMLR, pp. 6105-6114.

# Blockchain in the Quantum World

Arman Rasoodl Faridi, Faraz Masood, Ali Haider Thabet Shamsan, Mohammad Luqman, Monir Yahya Salmony

Department of Computer Science  
Aligarh Muslim University  
Aligarh, India

**Abstract**—Blockchain is one of the most discussed and highly accepted technologies, primarily due to its application in almost every field where third parties are needed for trust. Blockchain technology relies on distributed consensus for trust, which is accomplished using hash functions and public-key cryptography. Most of the cryptographic algorithms in use today are vulnerable to quantum attacks. In this work, a systematic literature review is done so that it can be repeated, starting with identifying the research questions. Focusing on these research questions, literature is analysed to find the answers to these questions. The survey is completed by answering the research questions and identification of the research gaps. It is found in the literature that 30% of the research solutions are applicable for the data layer, 24% for the application and presentation layer, 23% for the network layer, 16% for the consensus layer and only 1% for hardware and infrastructure layer. We also found that 6% of the solutions are not blockchain-based but present different distributed ledger technology.

**Keywords**—Blockchain; quantum computers; distributed ledger technology; security; systematic literature review; quantum attacks

## I. INTRODUCTION

Quantum computing is one of the latest technologies that has exploded in popularity in recent years. While the foundation of quantum mechanics has been more theoretical than practical for over 100 years, now the time has arrived when practically all firms are delving into it. In the late 1970s and early 1980s, research defining the fundamentals of quantum computing surfaced. Paul Benioff, an Argonne National Labs scientist, wrote a paper in 1979 that showed the theoretical foundation for quantum computing [1] and suggested that a quantum computer might be developed. Numerous businesses claim to be developing quantum computers, such as IBM, which is currently providing its clients with the first solutions in the form of a Quantum Gate Model. Google, Microsoft and many other companies are exploring similar machines.

Satoshi Nakamoto introduced the decentralised transfer and maintenance of digital assets that cannot be duplicated [2]. Distributed ledger technology (DLT) was initially used in finance, but it was subsequently discovered that it could be used whenever we desire to eradicate centralisation or intermediaries. The most widely used DLT is blockchain. There are other types of DLTs like IOTA [3], Hashgraph [4] etc., which are based on Directed Acyclic Graphs. Radix is also a DLT that uses a distributed database to store transactions [5]. Blockchain may be conceived as a sequence of interconnected blocks containing transactions. Every block stores the hash of the previous block, which results in a chain that is very difficult

to modify since modifying every transaction necessitates modifying the block, and modifying the block necessitates modifying the entire chain. Blockchain is the foundation of cryptocurrencies such as Bitcoin [2], Ethereum [6], Litecoin, etc.

Quantum computers cannot solve optimisation issues in a substantially scalable manner. In a universal infrastructure, there will be classical computers and quantum computers, with the quantum computer having a significant edge in terms of optimisation. Several quantum algorithms, such as Grover's algorithm [7], Shor's algorithm [8], and others, can solve some problems far quicker than conventional algorithms. Problems that have previously been almost insolvable will now be resolved in a reasonable period. In this regard, advancement in the quantum computing sector has piqued the curiosity of many researchers in both academia and industry.

Blockchain technology started to proliferate because of its nature to provide unbreakable data security, but once practical quantum computers are developed, they cannot provide such security [9]. Smart contracts can be hampered, and the whole technology will go down. The security of the blockchain is built on mathematical challenges that are extremely difficult for even the most powerful conventional computers to solve.

Public key cryptography protects cryptocurrencies. To breach public key encryption, quantum computers might possibly threaten the crypto industry, where some currencies are worth trillion of dollars. Encryption can be bypassed, allowing attackers to mimic legal owners of digital assets. All security assurances will be meaningless if quantum computing gets strong enough. To decrypt data, quantum computers will need thousands of qubits, compared to today's hundreds. Machines will also require persistent qubits that can do calculations for much longer than currently achievable.

NIST (National Institute of Standards and Technology) has already started finding, evaluating, and standardising public-key cryptography algorithms that are quantum-resistant [10]. However, it is necessary for the research community to primarily focus on blockchain technology. A lot of work is going on to create a quantum secure blockchain. To systematically analyse them following research questions are set:

RQ1: What challenges and security issues could occur due to the rise of quantum computers in blockchain technology?

RQ2: What are the various strategies and approaches used by researchers to make blockchain quantum resistant?

To answer these research questions, a systematic literature review has been undertaken. In Section 2, the research method is discussed in detail. Section 3 explores the basics of blockchain and quantum computing and the related challenges and solutions associated with these technologies. The survey results and answers to the research questions are discussed in Section 4, and the work is then concluded in Section 5 with future directions.

II. RESEARCH METHOD

This research utilised the SLR (Systematic Literature Review) method, as it helps to conduct the secondary research using a well-defined method. This approach gives us a framework to follow in order to discover, analyse, and evaluate relevant literature to find unbiased and reproducible answers to our research questions [11]. The parts of the process include planning, conducting and reporting on the review. Section 1 deals with the planning phase. Reporting is handled in Sections 3 and 4. This section goes through the phases of the review process, which includes:

A. Research Identification

This preliminary search aims to discover existing systematic reviews and determine the volume of studies that would be appropriate. A single search string is utilised instead of many search strings. Only databases related to the issue and widely accepted in the scholarly community are included. For this study, only IEEE Xplore, Elsevier, ACM Digital library, Springer Nature and Taylor & Francis is used.

The search string is formed to search throughout the metadata using the Boolean operator "AND," and the simple search term is ("Blockchain" AND "Quantum").

B. Inclusion and Exclusion Criteria

Inclusion and exclusion criteria should be based on the research questions to guarantee that the research questions can be effectively interpreted and that the studies are properly classified. Because the wide usage of blockchain grew in prominence after 2015, we chose all papers published after January 2016. We also limited the results to journal and conference articles, excluding online material, books, and magazines. If duplicate articles or corrections are found in any of these articles, they were removed. Finally, only articles written entirely in the English language are chosen.

C. Study Selection Process

We found 126 items in IEEE Xplore, 395 in Elsevier, 187 in the ACM digital library, 272 in Springer Nature, and 96 in Taylor & Francis Online using the specified search term and inclusion-exclusion criteria. A three-stage selection technique was implemented to guarantee that only relevant research articles were evaluated. Following the search, the results are extracted using keywords and titles. Following that, the abstracts of the papers were read, and the number of articles was decreased. Only high-quality studies that answered the research questions were picked in the last step, which involved reading whole articles and ranking them based on content. For efficient monitoring and control during the selection process, a separate folder was created for each evaluation stage, along with a new Excel sheet. The research is entirely transparent and

traceable as a result of this. Table I shows the step-by-step selection criteria, and Fig. 1 shows the count of publications that were included.

TABLE I. CRITERIA FOR ACCEPTANCE AT EACH STAGE

| Review Stage | Method                                                                | Criteria for acceptance                                                                                                         |
|--------------|-----------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| First        | Filter the articles using keywords and titles.                        | The title or keyword should be related to the research objective. Select the document for the next stage if there is any doubt. |
| Second       | Exclude articles based on abstracts                                   | Check if the abstract relates to the research question. In case of doubt, move the paper to the subsequent stage.               |
| Third        | Articles are excluded based on their entire text and article quality. | Papers that correspond to the research subject and proper experiments or mathematical proof is provided are selected.           |

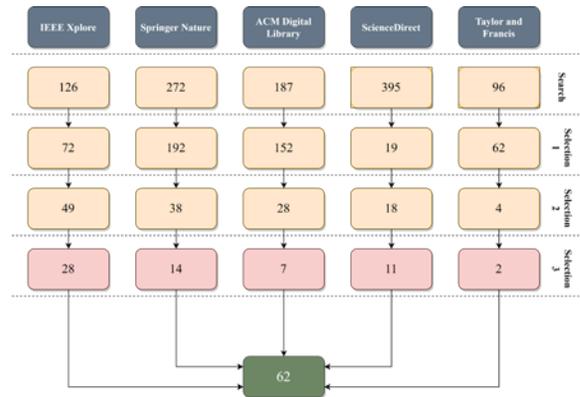


Fig. 1. Number of Papers Selected at each Stage.

D. Data Extraction

Once the analysis of the selected articles was completed, an excel file was created to record the data extracted from each publication. Table II shows the fields that were taken from each publication.

TABLE II. FIELDS USED FOR DATA EXTRACTION

| S. No | Field                            | Description                                                 |
|-------|----------------------------------|-------------------------------------------------------------|
| S1    | Title                            | The paper's title                                           |
| S2    | Database                         | Where an article is published                               |
| S3    | Rating                           | According to the content                                    |
| S4    | Experiment                       | Whether or not proper experimentation is carried out        |
| S5    | Mathematical Proof               | Whether or not mathematical proof is provided               |
| S6    | Architecture/Framework/Algorithm | Whether the architecture, framework, or algorithm is given. |
| S7    | Code                             | Whether source code is given to duplicate the results       |
| S8    | Survey                           | Is it a survey paper                                        |
| S9    | Problem identified               | Which type of issue is discussed in the paper               |
| S10   | Category of Solution             | What type of solution is provided                           |

### E. Data Synthesis

According to the research questions, all data taken from the selected publications was synthesised. This makes it simple to understand the challenges and different kinds of solutions provided.

As shown in Tables III and IV, a systematic data analysis assisted in the formalisation of specific categories related to the description of problems and solutions.

TABLE III. CHALLENGES BASED ON LAYERS

| S. No. | Layers                             | Articles                         |
|--------|------------------------------------|----------------------------------|
| 1      | Application and Presentation Layer | [17], [18], [27], [19]–[26]      |
| 2      | Consensus Layer                    | [28]–[36]                        |
| 3      | Network Layer                      | [37], [38], [47], [39]–[46]      |
| 4      | Data Layer                         | [48], [49], [58]–[65], [50]–[57] |
| 5      | Hardware and infrastructure layer  | [48]                             |
| 6      | Not based on Layers                | [66], [67]                       |

TABLE IV. SUMMARY OF SOLUTIONS FOUND IN THE LITERATURE

| S.No. | Solution                               | Article                                                                                 |
|-------|----------------------------------------|-----------------------------------------------------------------------------------------|
| 1     | Quantum Properties                     | [18], [21], [89], [26], [28], [31], [41], [44], [46], [70], [88]                        |
| 2     | Hash Based Signature                   | [24], [25], [50], [56], [58], [59], [71]                                                |
| 3     | Code-Based Cryptography                | [22]                                                                                    |
| 4     | Lattice Based Cryptography             | [20], [23], [53], [54], [57], [62], [63], [90], [38]–[40], [43], [45], [49], [51], [52] |
| 5     | Multivariate Cryptography              | [37], [55], [64]                                                                        |
| 6     | Directed Acyclic Graph                 | [66], [67]                                                                              |
| 7     | Quantum Blind Signature                | [18], [38], [42], [55], [70]                                                            |
| 8     | Quantum Walks                          | [61]                                                                                    |
| 9     | Hardware And Software Based Blockchain | [48]                                                                                    |
| 10    | Quantum Cloud Computing                | [17], [48]                                                                              |
| 11    | Post-Quantum Threshold Signature       | [29]                                                                                    |
| 12    | Quantum Random Oracle Model            | [43]                                                                                    |
| 13    | One Way Function                       | [60][65]                                                                                |
| 14    | Zero Knowledge Proof                   | [47][27]                                                                                |
| 16    | New Consensus                          | [21], [28]–[30], [32]–[36]                                                              |
| 17    | Review                                 | [9], [91]–[94]                                                                          |

Furthermore, as shown in Fig. 3 and 4, a frequency analysis is performed for the problems and solutions under study.

### III. SLR FINDINGS

We synthesised the data from the selected papers depending on the research questions. The problems are not clearly defined but presented as an overall solution to blockchain problems with quantum computing. In order to categorise them properly, the solutions provided are split based on different layers of blockchain. Every layer has different security requirements, so based on these layers, research articles are grouped. Also, some solutions are working of more than one layer, so these solutions are identified separately for each layer. First, we explain the problems in each layer and then different types of solutions studied in the literature.

#### A. Challenges and Issues

After the analysis, it is decided to represent blockchain in layers as shown in Fig. 2 and then understand the issues according to each layer. Dividing into layers make it easy to understand where research is still required. These layers, along with problems, are explained below:

1) *Hardware and infrastructure layer*: Internet users (peers) can now connect with other peers and share data as distributed systems are becoming more prevalent. This layer is responsible for creating virtual resources such as storage, networks, and servers. Nodes are the essential part of this layer because nodes are hardware devices that connect to the network and help make consensus in the blockchain. Infrastructure security frequently necessitates either limiting or prohibiting access to the node. So, improvement is needed at the infrastructure level to implement quantum blockchain properly.

2) *Data layer*: Data stored in blockchain depends on the type of blockchain-like Hyperledger Fabric [12] that contains channel information, whereas a Bitcoin blockchain needs to store the information about the sender, receiver, and amount. Blockchain network data is added only when consensus is reached among the nodes. Hash functions help in the easy identification of blocks and the detection of any changes made to the blocks. To ensure the confidentiality and integrity of the data stored on the blockchain, transactions are digitally signed. Blockchain uses asymmetric cryptography to secure information about the block, transactions, and transacting parties, among other things.

To sign a transaction, private keys are used, and anyone with the public key is used, and anyone with the public key can verify the signer. Because the encrypted data is also signed, digital signatures ensure data integrity. Every transaction in a block is hashed and organised in the form of a Merkle tree. In the Merkle tree hash of transactions are organised in the form of a binary tree. If any transaction is changed, then the whole Merkle tree is changed, which changes the whole block as the block contains the hash of the Merkle tree.



Fig. 2. Different Layers of the Blockchain.

As a result, any manipulation will render the signature invalid. Most blockchain systems depend significantly on a digital signature to improve security. These signatures rely on the difficulty of solving a mathematical problem, such as determining the factors of large integers. The data layer is highly dependent on these algorithms, and once practical quantum computers are developed, breaking these algorithms will be easy. As a result, this layer is too much vulnerable to quantum attacks.

3) *Network layer:* The network layer is in charge of inter-node communication and handles block propagation, transactions, and discovery. It is also called the propagation layer or peer to peer layer. In a peer-to-peer network, nodes share the workload to achieve a common goal in a distributed network. This layer ensures that nodes can discover other nodes in the network to interact, propagate, and synchronise information with other nodes. This layer also handles the propagation of the world state. A node can be a light node or a full node. Light nodes can merely retain the blockchain's header and send transactions. Full nodes are responsible for transaction verification and validation, mining, and consensus rule enforcement. They are in charge of ensuring the network's trustworthiness. So, it is needed that this layer uses quantum network in the future.

The term "Quantum Internet" [13]–[15] refers to the entire system, which comprises both quantum and classical packet switching networks. A traditional network in which hosts and routers can handle quantum information in the network graph structure is known as a quantum network. Between these nodes, there are classical channel for transferring classical data and quantum channel connections for transmitting quantum data.

4) *Consensus layer:* The rules that nodes follow to ensure that transactions are validated within those rules, and that blocks respect those rules is known as consensus. There is a consensus algorithm behind every blockchain as the trusted

third party is missing to validate transactions in case of conflict. The consensus layer is the most significant layer for any blockchain. Consensus protocols provide a set of irrefutable agreements between nodes in a distributed peer-to-peer network. Consensus keeps all of the nodes in sync. Consensus is in charge of validating the blocks, ordering them, and guaranteeing that everyone agrees. It is easy to attack this layer with the help of quantum computers. Attackers can search hash collisions, which can subsequently be used to change blocks in a network without impacting the integrity of the blockchain. Also, for mining, it is required to search a nonce, and with quantum computers, it will be very fast. This can enable an attacker to reconstruct the whole blockchain without getting detected by the network.

5) *Application layer:* The application layer includes smart contracts[16], chaincode[12], and decentralised apps (dApps). Smart contracts are digital contracts built on the blockchain that is automatically executed when particular events occur, or any external criterion are met. Chaincode is a collection of related smart contracts used to do a certain purpose. dApps are software applications that run on a blockchain network of computers rather than on a single device. Because they are decentralised, decentralised apps are free of the control and influence of a single authority.

dApps provide several advantages, including user privacy, developer independence and lack of censorship. The application layer comprises two layers: the application layer and the execution layer. The application layer is where end-users interact with the blockchain network, including scripts, APIs, user interfaces, and frameworks.

The execution layer, which includes smart contracts, underlying rules, and chaincode, is a sublayer. This sublayer contains the code and rules that are actually executed. A transaction is propagated from the application to the execution layer, but it is validated and executed by the semantic layer. The execution layer processes transactions and preserves the blockchain's deterministic nature. It receives instructions from the application layer. A smart contract code should not be pulled down or changed after being deployed on a blockchain, but it may be possible with quantum computers. Similarly, instead of making the smart contract more complex within the same technology, it is necessary that from now on, those researchers should start moving towards quantum-resistant smart contracts.

## B. Solutions

First, the basic concepts of quantum computing and some of the properties are discussed in this section. After that, the explanation of the solutions and methods that are found in the literature are discussed.

Quantum computing focuses on developing computer systems using quantum theory and quantum bits, or qubits. Quantum computers use subatomic particles' ability to exist in many states, i.e., it can be 0 or 1 simultaneously. Algorithms work by manipulating bits with gates, which change their states. The NAND gate is a universal gate, but the NAND gate's behaviour is not reversible because it accepts two inputs

and returns outputs that are not unique. In quantum computing, working with reversible gates is typically convenient since every reversible gate may be implemented on a quantum computer. The Toffoli gate is a reversible gate that takes three bits as input, can imitate the NAND gate.

Toffoli's gate converts  $(\alpha, \beta, \gamma)$  to  $(\alpha, \beta, (\gamma + \alpha * \beta) \bmod 2)$ . The Toffoli gate maps  $(\alpha, \beta, 1)$  to  $(\alpha, \beta, \alpha \text{ NAND } \beta)$  when  $\gamma = 1$ . Quantum computers are classically computationally ubiquitous because they can implement the Toffoli gate, even if the Toffoli gate alone is insufficient to implement any function on quantum states. The electron, which can have a spin pointing up or down, provides a simple physical prototype for this two-state system. These states are usually written as  $|0\rangle$  and  $|1\rangle$  in quantum mechanics as a convention. Quantum computers, unlike conventional computers, are not limited to manipulating only these two states. State superpositions, such as  $\frac{|0\rangle + |1\rangle}{\sqrt{2}}$  are also feasible. These two-state systems are known as quantum bits or qubits. Qubit states can alternatively be represented as two-dimensional vectors, for example.

$$|0\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, |1\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (1)$$

This is significant because multiplying the corresponding vectors allows gates to be represented mathematically as  $2 \times 2$  matrices acting on qubits.

A linear combination of  $|0\rangle$  and  $|1\rangle$  with complex coefficients can be used to describe the state  $|\psi\rangle$  of any given qubit, i.e.

$$|\psi\rangle = p|0\rangle + q|1\rangle, p, q \in \mathbb{C}. \quad (2)$$

A classical computer requires two complex numbers to describe an arbitrary quantum state; similarly, modelling  $n$  arbitrary quantum states on a classical computer requires  $2^n$  complex numbers and so a minimum of  $2^n$  Bits. By definition, a quantum computer requires just  $n$  qubits to describe  $n$  states. Modelling quantum systems on the classical computer will thus take the time that grows exponentially with the number of states  $n$ , whereas modelling the same system on a quantum computer only requires time that grows linearly with  $n$ . In other words, the classical computer takes  $O(2^n)$  time and the quantum computer takes  $O(n)$  time for this example.

The Hadamard gate is important in quantum computing. This gate, represented by  $H$ , has the following representations in matrix form and state notation:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \frac{1}{\sqrt{2}} \sum_{a,b \in \{0,1\}} (-1)^{ab} |a\rangle \langle b| \quad (3)$$

The Hadamard gate is a crucial component of quantum algorithms like Shor's algorithm, Grover's, and Simons algorithm as the Hadamard gate translates  $n$  qubits that are all in the same state to an equal superposition of the  $n$  qubits' potential states. Shor's algorithm, on the one hand, argues that, due to quantum mechanics, factorisation may be done in polynomial time rather than the exponential time, which is the basis of many public-key algorithms. Grover's algorithm, on the other hand, can cut the sufficient security strength of algorithms like the AES (Advanced Encryption Standard) in

half for a given key length, rendering infrastructures secured by them open to attack [7].

Shor's algorithm is noteworthy because it solves the complex problems of integer factorisation. The best extant algorithm for this problem is known as the generic number field sieve, and it operates in  $O(e^{(\log(N)^{1/3} \text{poly}(\log \log N))})$ , where  $\text{poly}$  is a complex polynomial. Shor's algorithm outperforms  $O((\log N)^3)$  in terms of speed.

Shor's approach employs the well-known Euclidean algorithm to compute the greatest common divisor (GCD) and then Simon's algorithm to gain the exponential speedup. Given an integer  $Z$ , one can compute  $\text{gcd}(f, Z)$  by selecting a random number  $f < Z$ . The problem is solved if  $f$  is a factor of  $Z$ ; otherwise,  $\text{gcd}(f, Z)$  must equal one. Assume  $M$  is the order of  $f$ , i.e., the lowest positive integer is  $M$  such that  $f^M \bmod N = 1$ . Then, as long as  $M$  is even and  $f^{M/2} + 1$  is not a multiple of  $Z$  (which is very likely), both  $\text{gcd}(f^{M/2} + 1, Z)$  and  $\text{gcd}(f^{M/2} - 1, Z)$  are factors of  $Z$ .

The problem is solved as long as the  $M$  matching to a particular  $f$  can be found. Consider the function  $g(x) = f^x \bmod Z$  to compute  $M$ . The task of computing  $M$  is thus reduced to period-finding for this function  $g$  since:

$$g(x + M) = f^{x+M} \bmod Z = f^M f^x \bmod Z = f^x \bmod Z = g(x) \quad (4)$$

The problem is solved by using Simon's algorithm. Simon's approach solves the period-finding problem, that is, calculating the period  $M$  of a function  $g$  that satisfies  $g(x) = g(x + M)$  for any  $x$ . This was a significant subproblem in Shor's method that provided an exponential speedup: Simon's algorithm runs in  $O(n)$  on a quantum computer and  $O(2^{n/2})$  on a classical computer [68].

With a uniform superposition of states over  $n$  qubits, Simon's algorithm computes the function  $g$  on the superposition, measures the answer, and applies the Hadamard gate to the  $n$  resultant qubit states. If the period  $M$  is represented as a  $n$ -bit vector  $\vec{M}$ , measuring the state after applying the Hadamard gate returns a vector orthogonal to  $\vec{M}$  with a high probability. After  $O(n)$  iterations of this process, one receives  $n - 1$  orthogonal vectors to  $\vec{M}$ . Because  $\vec{M}$  exists in an  $n$ -dimensional vector space, this is enough to determine the period  $M$ .

Grover's algorithm[69] is intended to tackle the problem of unstructured search. This problem can be described formally: given a function that transforms  $N$ -digit binary values to either 0 or 1, find  $x$ . Grover's algorithm is relatively straightforward to implement. To begin, use the Hadamard gate on a set of  $n$  qubits to generate a uniform superposition of states, where  $N = 2^n$ . Following that, a gate is built that rotates the uniform superposition towards the state  $|a\rangle$  corresponding to  $a$ . With a high probability, measuring the state after  $O(\sqrt{N})$  applications of this gate will yield  $|a\rangle$ . This is an improvement over the  $O(N)$  steps a random classical algorithm would take to find a best-case situation.

1) *Quantum properties*: A quantum state is a mathematical object that offers a probability distribution for each potential measurement of a system's outcomes. When we combine quantum states, we get another quantum state. Pure quantum states cannot be expressed as a mixture of other states, whereas mixed quantum states can be described as a combination of other states. Quantum computing performs computations by utilising the collective characteristics of quantum states, such as superposition, collapse, and entanglement.

A superposition of quantum states may be thought of as a linear combination of many quantum states, resulting in the development of a new valid quantum state. The basic states are  $|0\rangle$  and  $|1\rangle$ . All the Qubits are superposition on these basic states. Quantum superposition differs substantially from classical wave superposition. A superposition of  $2^m$  states, ranging from  $|0000\dots 0\rangle$  to  $|1111\dots 1\rangle$  will exist for a quantum computer with  $m$  qubits. The probability of a quantum state  $|\psi\rangle$  is  $|A_v|^2$  for any set of values  $v$  with probability amplitudes  $A_k \in \mathbb{C}^5$  in such a way that  $|\psi\rangle := \sum_v A_v |\psi_v\rangle$  for the measurement of  $|\psi\rangle$  resulting in  $\psi_v$ . Authors in [28] discussed the new consensus algorithm using quantum entanglement.

When one particle's quantum state cannot be characterised independently of the other particle's quantum state, they are said to be entangled. Even if the individual components are not in a defined state, the system's quantum state as a whole may be characterised. When two qubits become entangled, a one-of-a-kind relationship is established. The entanglement will be demonstrated by measurements, which may produce a value of 0 or 1 for individual qubits where the measurement of both the qubits will be the same. Even if the particles are separated by a significant distance, this is always true. For a quantum state  $|\psi\rangle$  with  $|\psi\rangle := \sum_v A_v |\psi_k^x, \psi_k^y\rangle$ , then on measurement of  $|\psi\rangle$  then probability X sees  $\psi_k^x$  and Y sees  $\psi_k^y$  is equal to 1.

While interacting with the outside environment, any wave function is reduced to a single eigenstate from the superposition of many eigenstates, and then it is called wave function collapse. In this case, the probability is 1 for all measurements of quantum state  $|\psi\rangle$  resulting in  $\psi_v$  where  $|\psi\rangle := \sum_v A_v |\psi_v\rangle$ , for some  $v$ .

A quantum channel can transfer both quantum and classical information. Quantum channels are trace-preserving mappings between spaces of fully positive operators. In other words, a quantum channel is just a quantum operation considered as a pipeline meant to transmit quantum information rather than simply the reduced dynamics of a system. Some solutions, as discussed in [18], [65] are based on quantum channels.

The idea of quantum key distribution (QKD) was initially presented in the 1970s, but it was not fully realised until the 1980s. QKD allows to sharing and distribute secret keys for cryptographic protocols. The essential thing is to keep them private, just between the communicating parties. Quantum superpositions or quantum entanglement and conveying information in quantum states may be used to develop a communication system that detects eavesdropping. If the extent of eavesdropping is less than a certain threshold, only then a

secure key can be generated otherwise, the communication is terminated. This is the general concept of Quantum cryptography that is why it is added as a property. Authors in [18], [19], [21], [26], [31], [41], [46], [70] discussed the usage of QKD for quantum blockchain.

2) *Hash-based signature*: The hash-based signature is used to utilise the cryptographic safe hash function properties. These properties include pre-image resistance, one-wayness and collision resistance. Hash-based signature systems rely entirely on the underlying safe cryptographic hash function, limiting the attack surface and cryptanalysis possibilities. By removing the need for several security components, hash-based signature systems substantially minimise implementation complexity. Any hash function that meets the security criteria of cryptographic hash functions can be employed to build hash-based signature algorithms. Because of this inherent flexibility, several underlying hash functions may be used to meet the required performance requirements based on the application-specific environment. Any difficult-to-invert function may be converted into a secure public-key signature system using hash-based cryptography. As a result, this might be a solution for post-quantum blockchains as discussed in [24], [32], [50], [71].

3) *Code-based cryptography*: All cryptosystems, symmetric or asymmetric, whose security is based, in part or entirely, on the difficulties of decoding a linear error-correcting code, perhaps chosen with some particular structure or in a particular family (for instance, quasi-cyclic codes, or Goppa codes) is code-based cryptography [72]. The ciphertext is a codeword with flaws that can only be corrected by the owner's private key (the Goppa code). Grover's algorithm does not significantly outperform earlier code-based cryptosystem attacks in terms of speed.

4) *Lattice-based cryptography*: A lattice is a collection of points having a periodic structure in  $n$ -dimensional space. Given  $n$ -linearly independent vectors  $v_1, v_2, v_3, \dots, v_n \in \mathbb{R}^m$  the set of vectors created by them is the lattice  $\mathcal{L}$

$$\mathcal{L}(v_1, v_2, v_3, \dots, v_n) = \{\sum x_i v_i \mid x_i \in \mathbb{Z}\} \quad (5)$$

A basis of the lattice is made up of the vectors  $v_1, v_2, v_3, \dots, v_n$ .

Because of its strong security proofs based on worst-case hardness, reasonably efficient implementations, and considerable simplicity, lattice-based cryptography [73] appears to promise post-quantum cryptography. In two ways, the worst-case security guarantee is critical. It helps us determine the cryptosystem's concrete parameters by ensuring that the cryptographic framework is free of fundamental flaws.

5) *Multivariate cryptography*: A set of (usually) quadratic polynomials over a finite field is a public map for a multivariate public-key cryptosystem (MPKC) [74]. In general, finding a solution to such structures is an NP-complete/-hard problem [75]. One of the intriguing instances is Patarin's Secret Fields [76], which generalises a suggestion by Matsumoto and Imai [77]. The NP-hardness of solving

nonlinear equations over a finite field underpins its fundamental security assumption. This is one of the most influential families of PKCs (public-key cryptography), as it can withstand even the most powerful quantum computers in the future. The MQPKC, unlike many other forms of PKC, cannot be solved quickly using Shor's algorithm with a conventional computer because it does not rely on any of the difficulties that Shor's algorithms can resolve.

6) *Directed acyclic graph*: A distributed ledger technology, a DAG [66], is an alternative to regular blockchain that seeks to solve blockchain technology's speed, scalability, and cost concerns. DAG is also a system that uses a digital ledger to keep track of transactions. DAG (Directed Acyclic Graph) is a more expressive outline than an entirely linear model. A DAG is a data or information structure that may be used to show a variety of difficulties. It is a topologically ordered acyclic graph. The node follows a specific sequence for each directed edge. Every DAG begins with a node with no parents and ends without children. There are no cyclic graphs on this page. A DAG is made up of nodes and arrows that connect them. By allowing many chains to exist on the system simultaneously, DAG can solve the single-chain problem of blockchain. IOTA is a DAG currency that is quite well-known. DAG Tangle is what they call it. It eliminates the need for miners in the verification process entirely. The white paper published by IOTA claims that Tangle is quantum-proof [3].

7) *Quantum blind signature*: A blind signature is a digital signature that blinds the message before it is signed. As a result, the message will go undetected by the signer. After that, the signed message will be unblinded. It functions as a standard digital signature and can be publicly verified. Blind signatures that can survive quantum attacks are referred to as "post-quantum blind signatures." Blind signatures have been widely used in the applications like the creation of e-cash and voting agreements. As a result, new quantum blind signature technologies will be necessary for the future. This solution works with other solutions like lattice-based or multivariate cryptography in order to provide quantum-resistant blockchain [18], [19], [39], [42], [54], [55].

8) *Quantum walks*: A random walk is a random process in mathematical space that defines a path consisting of a series of random steps, as defined by Pearson in 1905 [78]. Random walks are essential in solving practical issues since they can be used to evaluate and mimic the unpredictability of items and determine the correlation between them. Quantum walks were introduced in 1993 [79]. The polar opposite of traditional random walks is quantum walks. Quantum walks differ from regular random walks in that they do not converge to any limiting distributions and are much faster because of Quantum interference [79]–[81]. Quantum walks can outperform any traditional algorithm by order of magnitude. The two types of quantum walk-based algorithms are continuous time-based and discrete time-based algorithms [82].

9) *Hardware and software based blockchain*: As shown, blockchain implementation may be implemented into many different technology stack layers. So, hardware-based security is also essential. It may involve hardware-based secure key storage or hardware replacement for quantum channels. Hardware-based key storage is already being developed as cold wallets, but it must also be quantum secure. The authors in [48] develop a quantum computing device as a multi-input multi-output quantum channel.

10) *Quantum cloud computing*: In a cloud computing environment, a cloud quantum computer is a computer that can be accessed over the internet. Users may now make use of a variety of cloud quantum computing services to solve complicated issues that demand a lot of computational power. The design and performance of different cloud quantum computing systems vary. Solutions discussed in [17], [48] used quantum cloud computing.

11) *Post-quantum threshold signature*: Threshold signature [83] is a unique digital signature that can be used to identify a group of users. It is generated by an authorised subset of the private keys. The public keys are already generated with these private keys. It is very easy to verify these signatures as only a single public key and a single signature is enough. If at least  $n$  users out of  $m$  users efficiently sign the message, then the system is known as  $(n, m)$  threshold. The solution discussed in [29] is based on solving quadratic equations in a finite field, an NP-hard problem. This system is a threshold signature system and is considered safe even after developing a powerful quantum computer.

12) *Quantum random oracle model*: In a random oracle [84], anyone may give it an input and output of fixed length. If someone has already requested the input, the oracle will provide the identical result. If the oracle receives an input that it has not seen before, it generates a random output. To make the whole system secure, it is needed to replace all the hash functions used in the system with random oracles. Traditional oracle models can be easily attacked by using quantum superposition. This may result in the failure of many classical security proofs and must be rewritten. Quantum random oracle along with lattice-based solutions are discussed in [43].

13) *One way function*: A one-way function is easy to compute on all inputs but complex to invert given the image of a random input. In many cryptographic systems, one-way functions have proven useful primitive. Extensive work on one-way quantum functions has also been done in the post-quantum period. These one-way functions accept outputs of the quantum states by taking classical bit strings as input. Many information-theoretically secured digital signature techniques rely on the one-wayness characteristic of these functions [85], [86] have been proposed. To authenticate both classical bit strings and quantum states, these one-way functions should be both quantum-classical and classical-quantum in design. As a result, [60] developed quantum

money systems based exclusively on the security of one-way functions that are resistant to quantum attacks.

14)Zero-knowledge proof: Zero-Knowledge Proofs (ZKPs) enable data to be validated without disclosing the data itself. As a result, they have the potential to transform the way data is gathered, used, and transacted. Each transaction is assigned a 'verifier' and a 'prover'. In a ZKPs transaction, the prover tries to prove something to the verifier without revealing anything about it. The authors of [47] suggest employing two indistinguishable hash functions combined with ZKPs protocols to ensure security against quantum attacks.

#### IV. RESULT AND DISCUSSION

Blockchain is an up-and-coming technology, and it is assumed that it is the foundation of web 3.0. Quantum computing is not just theoretical now, as can be seen with the development of quantum computers by google (72 qubits), Xanadu (24 qubits), IBM (127 qubits), Intel (49 qubits) etc. Quantum Computers are real threats to blockchain technology, as discussed in the article. Our literature review found that to make blockchain stable even with quantum computers, work must be done at all the layers of blockchain, not just one layer. By that, we can genuinely make a quantum-resistant blockchain.

The focus was mainly on the research questions in the survey, and both the research questions were answered. The security threats on the blockchain are divided based on the layers of the blockchain and based on that we analysed the papers. As shown in Fig 3, most of the work (i.e., 30%) mainly focused on the data layer, which seems likely because mainly encryption and transactions are handled in this layer. The next area of focus was the application and presentation layer, with 24% of articles has shown the work on that. This layer includes applications based on blockchain, which may include smart contracts or chain codes. Therefore, the security of this layer is essential; however, the focus of the articles found concerning specific applications, so the focus should be on general solutions as well. For the network layer, it is found that 23% of papers work to find secure quantum networking and 16% of the articles found work on either changing the consensus algorithm or proposing the new algorithm in itself. Only 1% of articles discussed infrastructure and 6% about working with distributed ledger other than blockchain like IOTA, which is based on the directed acyclic graph.

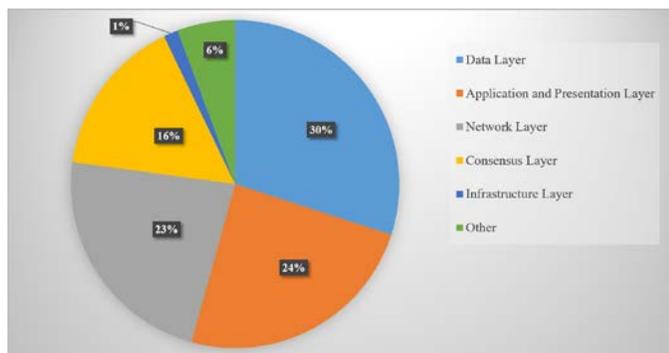


Fig. 3. Security Challenges Identified based on Blockchain Layers.

Next, we aimed to categorise solutions based on four categories only, i.e. Hash-based signatures, Code-based cryptography, lattice-based cryptography and Multivariate cryptography . However, instead of sticking to these four, we decided to make it more transparent and focus on the essential solutions. As shown in Fig. 4, around 25% of papers focused on lattice-based cryptography.

Consensus is necessary for blockchain for the settlement of the transaction. 14% of the papers proposed a new or modified consensus algorithm using either a new hash function, digital signature, or quantum properties. As hash functions and digital signatures are the backbones of blockchain technology, it is necessary to create new or modified signature schemes, and it has been found that 11% of research papers focused on hash-based signatures and 11% of the paper focused on quantum blind signatures. So these are the key areas where research is going on. Analysing the problems and solutions, it is clear that some layers still need some work, like the infrastructure and consensus layers. These layers are also necessary. Findings also suggest that some authors give a solution for one layer and claim that the blockchain will be posted quantum blockchain To make blockchain safe from quantum attacks, it is necessary to create the solution keeping in mind all the layers and find a solution that covers the problems of each layer.

This paper mainly focused on the research found in the literature to increase blockchain security in the post-quantum era. Some literature having reviews based on different focus areas are also found, like authors in [95] focused on proof of stake only, authors in [91] discuss the survival of DLTs after quantum computing. However, it was not thorough, focus on bridging quantum, and classical computing is done in [9], authors in [10] has done a good survey on post-quantum blockchain and compared the significant features of the post-quantum encryption cryptosystems that advanced to the second round of NIST call. This study is restricted for database and year selection to make the review process repeatable and free from any bias.

Even after that, many articles are obtained, evaluated, read, classified, and summarised, and answers to the research questions are presented. The findings suggest that it is needed to see blockchain systems in layers, and researchers should provide solutions to the quantum attacks based on these layers.

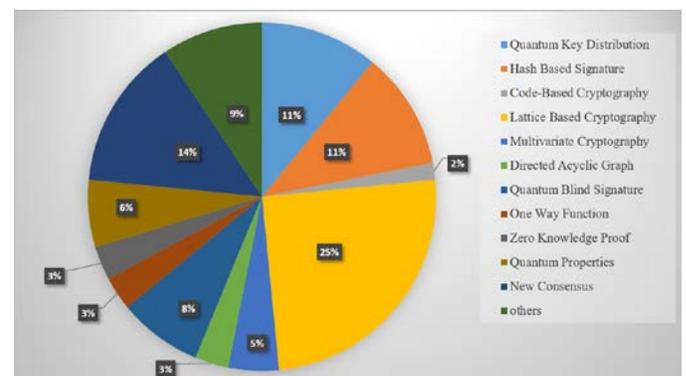


Fig. 4. Frequency Analysis of Solution for the Discussed Challenges.

## V. CONCLUSION

This study starts with the development of the research questions. To appropriately answer these questions, the systematic literature review is done, and the process is explained in-depth, including the database selection, search process, inclusion and exclusion criterion, creating and extraction of fields and summarising the results. During this process, we found and classified threats based on blockchain layers. Some of the threats were spread over different layers, so these threats are discussed individually for a proper explanation. Many different solutions are also found regarding these threats. The mapping between these threats and solutions has been presented, keeping in mind the full proof solution of post-quantum blockchain.

We discovered that blockchain could operate after quantum computers, but it must work on every layer of the blockchain network, or the solution will not be feasible. Even after developing solutions, they must be thoroughly tested in the real world. If a new application, whether decentralised or not, is being created on the blockchain, quantum attacks should be considered from the planning phase. It has been discovered that blockchain in its current form is unsuitable and must be modified. In the future, researchers will need to create similar solutions and test them for all such issues that have yet to be solved or discussed.

## REFERENCES

- [1] P. Benioff, "The computer as a physical system: A microscopic quantum mechanical Hamiltonian model of computers as represented by Turing machines," *J. Stat. Phys.*, vol. 22, no. 5, pp. 563–591, 1980.
- [2] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System | Satoshi Nakamoto Institute," 2008.
- [3] M. Divya and N. Biradar, "IOTA-Next Generation Block chain," *Int. J. Eng. Comput. Sci.*, vol. 7, pp. 23823–23826, 2018.
- [4] L. Baird, M. Harmon, and P. Madsen, "Hedera: A governing council and public hashgraph network - The trust layer of the internet," *Whitepaper*, pp. 1–27, 2018.
- [5] D. Hughes, "Radix-tempo," *Radix DTL Whitepaper*, 2017.
- [6] G. Wood, "ETHEREUM: A SECURE DECENTRALISED GENERALISED TRANSACTION LEDGER BYZANTIUM VERSION e94bda," 2018.
- [7] J. Mulholland, M. Mosca, and J. Braun, "The Day the Cryptography Dies," *IEEE Secur. Priv.*, vol. 15, no. 4, pp. 14–21, 2017.
- [8] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring," in *Proceedings 35th Annual Symposium on Foundations of Computer Science, 1994*, pp. 124–134.
- [9] F. Glover, G. Kochenberger, M. Ma, and Y. Du, "Quantum Bridge Analytics II: QUBO-Plus, network optimization and combinatorial chaining for asset exchange," *4or*, vol. 18, no. 4, pp. 387–417, 2020.
- [10] T. M. Fernandez-Carames and P. Fraga-Lamas, "Towards Post-Quantum Blockchain: A Review on Blockchain Cryptography Resistant to Quantum Computing Attacks," *IEEE Access*, vol. 8, pp. 21091–21116, 2020.
- [11] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," 2007.
- [12] E. Androulaki et al., "Hyperledger Fabric: A Distributed Operating System for Permissioned Blockchains," in *Proceedings of the Thirteenth EuroSys Conference, 2018*, p. 15.
- [13] H. J. Kimble, "The quantum internet," *Nature*, vol. 453, no. 7198, pp. 1023–1030, 2008.
- [14] L. Gyongyosi and S. Imre, "Entanglement-Gradient Routing for Quantum Networks," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, 2017.
- [15] L. Gyongyosi, S. Imre, and H. V. Nguyen, "A Survey on Quantum Channel Capacities," *IEEE Commun. Surv. Tutorials*, vol. 20, no. 2, pp. 1149–1205, 2018.
- [16] A. Saveliev, "Contract Law 2.0: Smart Contracts As the Beginning of the End of Classic Contract Law," *SSRN Electron. J.*, 2017.
- [17] W. Dai, "Platform modelling and scheduling game with multiple intelligent cloud-computing pools for big data," *Math. Comput. Model. Dyn. Syst.*, vol. 24, no. 5, pp. 506–552, 2018.
- [18] Z. Cai, J. Qu, P. Liu, and J. Yu, "A blockchain smart contract based on light-weighted quantum blind signature," *IEEE Access*, vol. 7, pp. 138657–138668, 2019.
- [19] J. L. Zhang, M. S. Hu, Z. J. Jia, Bei-Gong, and L. P. Wang, "A Novel E-payment Protocol Implemented by Blockchain and Quantum Signature," *Int. J. Theor. Phys.*, vol. 58, no. 4, pp. 1315–1325, 2019.
- [20] A. H. Karbasi and S. Shahpasand, "A post-quantum end-to-end encryption over smart contract-based blockchain for defeating man-in-the-middle and interception attacks," *Peer-to-Peer Netw. Appl.*, vol. 13, no. 5, pp. 1423–1441, 2020.
- [21] X. Sun, Q. Wang, P. Kulicki, and M. Sopek, "A Simple Voting Protocol on Quantum Blockchain," *Int. J. Theor. Phys.*, vol. 58, no. 1, pp. 275–281, 2019.
- [22] S. Gao, D. Zheng, R. Guo, C. Jing, and C. Hu, "An anti-quantum e-voting protocol in blockchain with audit function," *IEEE Access*, vol. 7, pp. 115304–115316, 2019.
- [23] Y. Lee, B. Son, H. Jang, J. Byun, T. Yoon, and J. Lee, "Atomic cross-chain settlement model for central banks digital currency," *Inf. Sci. (Ny.)*, vol. 580, pp. 838–856, 2021.
- [24] S. Suhail, R. Hussain, A. Khan, and C. S. Hong, "On the Role of Hash-Based Signatures in Quantum-Safe Internet of Things: Current Solutions and Future Directions," *IEEE Internet Things J.*, vol. 8, no. 1, pp. 1–17, 2021.
- [25] R. Amos, M. Georgiou, A. Kiayias, and M. Zhandry, "One-shot signatures and applications to hybrid quantum/classical authentication," *Proc. Annu. ACM Symp. Theory Comput.*, pp. 255–268, 2020.
- [26] H. Abulkasim, A. Mashatan, and S. Ghose, "Quantum-based privacy-preserving sealed-bid auction on the blockchain," *Optik (Stuttg.)*, vol. 242, no. April, p. 167039, 2021.
- [27] S. Dolev and Z. Wang, "SodsMPC: FSM based Anonymous and Private Quantum-safe Smart Contracts," *2020 IEEE 19th Int. Symp. Netw. Comput. Appl. NCA 2020*, 2020.
- [28] Y. L. Gao, X. B. Chen, G. Xu, K. G. Yuan, W. Liu, and Y. X. Yang, "A novel quantum blockchain scheme base on quantum entanglement and DPoS," *Quantum Inf. Process.*, vol. 19, no. 12, pp. 1–15, 2020.
- [29] H. Yi, Y. Li, M. Wang, Z. Yan, and Z. Nie, "An Efficient Blockchain Consensus Algorithm Based on Post-Quantum Threshold Signature," *Big Data Res.*, vol. 26, p. 100268, 2021.
- [30] J. Wang et al., "GSCS: General Secure Consensus Scheme for Decentralized Blockchain Systems," *IEEE Access*, vol. 8, pp. 125826–125848, 2020.
- [31] G. Iovane, "MuReQua Chain: Multiscale Relativistic Quantum Blockchain," *IEEE Access*, vol. 9, pp. 39827–39838, 2021.
- [32] J. Chen, W. Gan, M. Hu, and C. M. Chen, "On the construction of a post-quantum blockchain for smart city," *J. Inf. Secur. Appl.*, vol. 58, no. March, p. 102780, 2021.
- [33] J. Chen, W. Gan, M. Hu, and C.-M. Chen, "On the Construction of a Post-Quantum Blockchain," in *2021 IEEE Conference on Dependable and Secure Computing (DSC)*, 2021, vol. 11, no. 2, pp. 1–8.
- [34] J. Seet and P. Griffin, "Quantum Consensus," in *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2019*, 2019, vol. 0.
- [35] F. M. Ablayev, D. A. Bulychkov, D. A. Sapaev, A. V. Vasiliev, and M. T. Ziatdinov, "Quantum-Assisted Blockchain," *Lobachevskii J. Math.*, vol. 39, no. 7, pp. 957–960, 2018.
- [36] S. Dolev and Z. Wang, "SodsBC: Stream of Distributed Secrets for Quantum-safe Blockchain," *Proc. - 2020 IEEE Int. Conf. Blockchain, Blockchain 2020*, pp. 247–256, 2020.

- [37] S. B. Far and M. R. Asaar, "A blockchain-based quantum-secure reporting protocol," no. April, 2021.
- [38] C. Y. Li, X. B. Chen, Y. L. Chen, Y. Y. Hou, and J. Li, "A New Lattice-Based Signature Scheme in Post-Quantum Blockchain Network," *IEEE Access*, vol. 7, pp. 2026–2033, 2019.
- [39] P. Zhang, H. Jiang, Z. Zheng, P. Hu, and Q. Xu, "A New Post-Quantum Blind Signature from Lattice Assumptions," *IEEE Access*, vol. 6, pp. 27251–27258, 2018.
- [40] Q. Li, M. Luo, C. Hsu, L. Wang, and D. He, "A Quantum Secure and Noninteractive Identity-Based Aggregate Signature Protocol From Lattices," *IEEE Syst. J.*, no. Id, pp. 1–11, 2021.
- [41] M. T. Azhar, M. B. Khan, and A. U. R. Khan, "Blockchain based secure crypto-currency system with quantum key distribution protocol," 2019 8th Int. Conf. Inf. Commun. Technol. ICICT 2019, pp. 31–35, 2019.
- [42] M. Bhavin, S. Tanwar, N. Sharma, S. Tyagi, and N. Kumar, "Blockchain and quantum blind signature-based hybrid scheme for healthcare 5.0 applications," *J. Inf. Secur. Appl.*, vol. 56, no. December 2020, p. 102673, 2021.
- [43] N. Alkeilani Alkadri et al., "Deterministic Wallets in a Quantum World," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 1017–1031, 2020.
- [44] Q. Zhou and H. Lv, "Multi-secret Sharing Model based on Hermite Interpolation Polynomial and Quantum Graph State," *Int. J. Theor. Phys.*, vol. 59, no. 8, pp. 2271–2293, 2020.
- [45] C. Ma and M. Jiang, "Practical Lattice-Based Multisignature Schemes for Blockchains," *IEEE Access*, vol. 7, pp. 179765–179778, 2019.
- [46] H. Chen, "Quantum relay blockchain and its applications in key service," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, pp. 95–99, 2020.
- [47] M. Hari Krishnan and K. V. Lakshmy, "Secure Digital Service Payments using Zero Knowledge Proof in Distributed Network," 2019 5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019, pp. 307–312, 2019.
- [48] W. Dai, "Quantum-computing with AI & blockchain: modelling, fault tolerance and capacity scheduling," *Math. Comput. Model. Dyn. Syst.*, vol. 25, no. 6, pp. 523–559, 2019.
- [49] R. Saha et al., "A Blockchain Framework in Post-Quantum Decentralization," *IEEE Trans. Serv. Comput.*, vol. 11, no. 4, 2021.
- [50] A. Shafarenko, "A PLS blockchain for IoT applications: protocols and architecture," *Cybersecurity*, vol. 4, no. 1, 2021.
- [51] Y. L. Gao, X. B. Chen, Y. L. Chen, Y. Sun, X. X. Niu, and Y. X. Yang, "A Secure Cryptocurrency Scheme Based on Post-Quantum Blockchain," *IEEE Access*, vol. 6, no. Part II, pp. 27205–27213, 2018.
- [52] J. Di, T. Xie, S. Fan, W. Jia, and S. Fu, "An Anti-Quantum Signature Scheme over Ideal Lattice in Blockchain," *Proc. - 2020 Int. Symp. Comput. Eng. Intell. Commun. ISCEIC 2020*, pp. 218–226, 2020.
- [53] W. Yin, Q. Wen, W. Li, H. Zhang, and Z. Jin, "An anti-quantum transaction authentication approach in blockchain," *IEEE Access*, vol. 6, pp. 5393–5401, 2017.
- [54] C. Li, Y. Tian, X. Chen, and J. Li, "An efficient anti-quantum lattice-based blind signature for blockchain-enabled systems," *Inf. Sci. (Ny.)*, vol. 546, pp. 253–264, 2021.
- [55] H. Yi, "A traceability method of biofuel production and utilization based on blockchain," *Fuel*, no. October, p. 122350, 2021.
- [56] L. A. Lizama-Perez, "Digital signatures over hash-entangled chains," *SN Appl. Sci.*, vol. 1, no. 12, pp. 1–8, 2019.
- [57] M. F. Esgin, R. K. Zhao, R. Steinfeld, J. K. Liu, and D. Liu, "Matrix: Efficient, scalable and post-quantum blockchain confidential transactions protocol," *Proc. ACM Conf. Comput. Commun. Secur.*, pp. 567–584, 2019.
- [58] F. Shahid, I. Ahmad, M. Imran, and M. Shoaib, "Novel One Time Signatures (NOTS): a Compact Post-Quantum Digital Signature Scheme," *IEEE Access*, vol. 8, pp. 15895–15906, 2020.
- [59] J. Alupotha and X. Boyen, "Origami Store: UC-Secure Foldable Datachains for The Quantum Era," *IEEE Access*, vol. 9, 2021.
- [60] A. Behera and G. Paul, "Quantum to classical one-way function and its applications in quantum money authentication," *Quantum Inf. Process.*, vol. 17, no. 8, pp. 1–24, 2018.
- [61] A. A. Abd El-Latif, B. Abd-El-Atty, I. Mehmood, K. Muhammad, S. E. Venegas-Andraca, and J. Peng, "Quantum-Inspired Blockchain-Based Cybersecurity: Securing Smart Edge Utilities in IoT-Based Smart Cities," *Inf. Process. Manag.*, vol. 58, no. 4, p. 102549, 2021.
- [62] C. Wu, L. Ke, and Y. Du, "Quantum resistant key-exposure free chameleon hash and applications in redactable blockchain," *Inf. Sci. (Ny.)*, vol. 548, pp. 438–449, 2021.
- [63] M. Kansal and R. Dutta, Round optimal secure multisignature schemes from lattice with public key aggregation and signature compression, vol. 12174 LNCS. Springer International Publishing, 2020.
- [64] H. Yi, "Secure Social Internet of Things Based on Post-Quantum Blockchain," *IEEE Trans. Netw. Sci. Eng.*, vol. 4697, no. c, pp. 1–8, 2021.
- [65] S. Singh, N. K. Rajput, V. K. Rathi, H. M. Pandey, A. K. Jaiswal, and P. Tiwari, "Securing Blockchain Transactions Using Quantum Teleportation and Quantum Digital Signature," *Neural Process. Lett.*, 2020.
- [66] I. Keidar, E. Kokoris-Kogias, O. Naor, and A. Spiegelman, "All You Need is DAG," *Proc. Annu. ACM Symp. Princ. Distrib. Comput.*, pp. 165–175, 2021.
- [67] S. Suhail, R. Hussain, A. Khan, and C. S. Hong, "Orchestrating product provenance story: When IOTA ecosystem meets electronics supply chain space," *Comput. Ind.*, vol. 123, p. 103334, 2020.
- [68] D. Bacon, "Quantum Computing Simon's Algorithm," *Lect. Notes, Univ. Washingt.*, pp. 1–5, 2006.
- [69] L. K. Grover, "A Fast Quantum Mechanical Algorithm for Database Search," in *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, 1996, pp. 212–219.
- [70] X. Lin Gou, R. Hua Shi, W. Gao, and M. Wu, "A novel quantum E-payment protocol based on blockchain," *Quantum Inf. Process.*, vol. 20, no. 5, pp. 1–17, 2021.
- [71] E. Santoso and A. M. Barmawi, "Improving the Performance of Blockchain Based Digital Contract Using Niederreiter Method," *PervasiveHealth Pervasive Comput. Technol. Healthc.*, pp. 19–26, 2020.
- [72] R. J. McEliece, "A Public-Key Cryptosystem Based On Algebraic Coding Theory," *The Deep Space Network Progress Report*, vol. 42, no. 44, pp. 114–116, 1978.
- [73] M. Ajtai and C. Dwork, "A Public-Key Cryptosystem with Worst-Case/Average-Case Equivalence," in *Proceedings of the Twenty-Ninth Annual ACM Symposium on Theory of Computing*, 1997, pp. 284–293.
- [74] J. Ding and B.-Y. Yang, "Multivariate public key cryptography," in *Post-quantum cryptography*, Springer, 2009, pp. 193–241.
- [75] J. Ding and A. Petzoldt, "Current State of Multivariate Cryptography," *IEEE Secur. Priv.*, vol. 15, no. 4, pp. 28–36, 2017.
- [76] J. Patarin, "Hidden fields equations (HFE) and isomorphisms of polynomials (IP): Two new families of asymmetric algorithms," in *International Conference on the Theory and Applications of Cryptographic Techniques*, 1996, pp. 33–48.
- [77] J. Patarin, "Cryptanalysis of the Matsumoto and Imai Public Key Scheme of Eurocrypt'98," *Des. Codes Cryptogr.*, vol. 20, no. 2, pp. 175–209, 2000.
- [78] K. PEARSON, "The Problem of the Random Walk," *Nature*, vol. 72, no. 1865, p. 294, 1905.
- [79] Y. Aharonov, L. Davidovich, and N. Zagury, "Quantum random walks," *Phys. Rev. A*, vol. 48, pp. 1687–1690, 1993.
- [80] E. Farhi and S. Gutmann, "Quantum computation and decision trees," *Phys. Rev. A*, vol. 58, no. 2, pp. 915–928, Aug. 1998.
- [81] A. M. Childs, R. Cleve, E. Deotto, E. Farhi, S. Gutmann, and D. A. Spielman, "Exponential Algorithmic Speedup by a Quantum Walk," in *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, 2003, pp. 59–68.
- [82] J. Kempe, "Quantum random walks: An introductory overview," *Contemp. Phys.*, vol. 44, no. 4, pp. 307–327, Jul. 2003.
- [83] G. Bleumer, "Threshold Signature," in *Encyclopedia of Cryptography and Security*, H. C. A. van Tilborg, Ed. Boston, MA: Springer US, 2005, pp. 611–614.

- [84] M. Bellare and P. Rogaway, "Random Oracles Are Practical: A Paradigm for Designing Efficient Protocols," in *Proceedings of the 1st ACM Conference on Computer and Communications Security*, 1993, pp. 62–73.
- [85] D. Gottesman and I. Chuang, "Quantum Digital Signatures," 2001.
- [86] X. Lü and D. Feng, "Quantum digital signature based on quantum one-way functions," *7th Int. Conf. Adv. Commun. Technol. 2005, ICACT 2005.*, vol. 1, pp. 514–517, 2005.
- [87] J. Buchmann and J. Ding, *Post-Quantum Cryptography: Second International Workshop, PQCrypto 2008 Cincinnati, OH, USA October 17-19, 2008 Proceedings*, vol. 5299. Springer Science & Business Media, 2008.
- [88] K. Narendra and G. Aghila, "Fortis-ámyna-smart contract model for cross border financial transactions," *ICT Express*, vol. 7, no. 3, pp. 269–273, 2021.
- [89] A. Ahuja, "TensorFlip: A Fast Fully-Decentralized Computational Lottery for Cryptocurrency Networks," in *2021 International Conference on COMMunication Systems and NETWORKS, COMSNETS 2021*, 2021, vol. 2061, pp. 246–253.
- [90] H. Karim and D. B. Rawat, "TollsOnly Please &#x2013; Homomorphic Encryption for Toll Transponder Privacy in Internet of Vehicles," *IEEE Internet Things J.*, vol. 4662, no. c, 2021.
- [91] Q. Zhu, S. W. Loke, R. Trujillo-Rasua, F. Jiang, and Y. Xiang, "Applications of distributed ledger technologies to the internet of things: A survey," *ACM Comput. Surv.*, vol. 52, no. 6, 2019.
- [92] E. Giusto, M. G. Vakili, F. Gandino, C. Demartini, and B. Montrucchio, "Quantum Pliers Cutting the Blockchain," *IT Prof.*, vol. 22, no. 6, pp. 90–96, 2020.
- [93] W. Ma, W. J. Chen, and W. Paweenbampen, "Survey of Whether Blockchain Can Replace Other Online-Payment," *2019 2nd IEEE Int. Conf. Hot Information-Centric Networking, HotICN 2019*, pp. 84–89, 2019.
- [94] R. Koch and M. Golling, "The cyber decade: Cyber defence at a X-ing point," *Int. Conf. Cyber Conflict, CYCON*, vol. 2018-May, pp. 159–185, 2018.
- [95] A. M. Khalifa, A. M. Bahaa-Eldin, and M. A. Sobh, "Quantum attacks and defenses for proof-of-stake," *Proc. - ICCES 2019 2019 14th Int. Conf. Comput. Eng. Syst.*, pp. 112–117, 2019.

# Design and Implementation of Deep Depth Decision Algorithm for Complexity Reduction in High Efficiency Video Coding (HEVC)

Helen K Joy<sup>1</sup>, Manjunath R Kounte<sup>2</sup>  
School of ECE  
REVA University  
Bengaluru, India

B K Sujatha<sup>3</sup>  
Department of Electronics and Telecommunication  
Ramaiah Institute of Technology  
Bengaluru, India

**Abstract**—High efficiency video (HEVC) coding made its mark as a codec which compress with low bit rate than its preceding codec that is H.264, but the factor that stop HEVC from many applications is its complex encoding procedure. The rate distortion optimisation (RDO) cost calculation in HEVC consume complex calculations. In this paper, we propose a method to cross out the issue of complex calculations by replacing the traditional inter-prediction procedure of brute force search for RDO by a deep convolutional neural network to predict and perform this process. In the first step, the modelling of the deep depth decision algorithm is done with optimum specifications using convolutional neural network (CNN). In the next step, the model is designed and trained with dataset and validated. The trained model is tested by pipelining it to the original HEVC encoder to check its performance. We also evaluate the efficiency of the model by comparing the average time of encoding for various resolution video input. The testing is done with mutually independent input to maintain the accuracy of the system. The system shows a substantial saving in encoding time that proves the complexity reduction in HEVC.

**Keywords**—CNN; HEVC; deep learning; RDO; encoding time; complexity reduction

## I. INTRODUCTION

Video compression is an area to explore while considering the flourish of video acquisition devices, social media, live transfer of videos etc. The high efficiency video encoding HEVC system possess a better compression compared to its previous system advanced video coding AVC. However, the computational complexity of HEVC is a matter of discussion because of its Rate distortion optimisation [1] (RDO) cost calculation for coding tree unit [2] (CTU). There for this computational complexity in HEVC is a matter of research interest, the focus will be to reduce the computational complexity[3] with better efficiency. Before going into the details let's review the evolution of HEVC its drawback and goodness compared to its ancestral system.

ITU-T video compression standard introduced H.261 is the year November 1988. In the H.26x family, first member, H.261 in video coding standards in the domain of the VCEG (ITU-T Video Coding Experts Group) then Specialists Group on Coding for Visual Telephony [4]. H.261 was originally

designed to transmit data over ISDN lines with data rates as multiples of 64 Kbit/s. The coding algorithm was designed in such a way to work at video bit rates in 40 Kbit/s to 2 Mbit/s [5]. MPEG-2 consists of “three different kinds of coded frames: I-frame /intra-coded frames, P -frame/predictive-coded frames, and B-frame/ bidirectionally-predictive-coded frames” [3]. The I-frame is a single uncompressed or raw frame that is a separately-compressed version. “The I-frame coding takes the advantage of spatial redundancy and the persistence of vision of human eye ie the inability of human eye to detect several changes in the image. I-frames do not depend on data in the previous or the next frames,”[6] Unlike in P-frames and B-frames, and because of that its coding matches with the still photography. The raw frame is spitted into 8X8-pixel blocks. The data in all block is transformed using discrete cosine transform (DCT)[6]and results is a matrix of size 8x8 of coefficients that have real number values. The DCT transform converts spatial domain into frequency domain, but it does not change the information in the block; if the DCT is calculated with perfect precision, the original block can be recovered clearly by applying the inverse discreet cosine transform [5] “H.263 [7]is a popular video compression standard for low-bit-rate compressed format focusing on videoconferencing. It was standardized by the organisation ITU-T, Video Coding Experts Group (VCEG) in 1995/1996”[4] . H.263 is the member of the H.26x family of video coding standards in the domain, ITU-T. Like other H.26x standards, H.263 is also based on (DCT) discrete cosine transform video compression. It was later advanced to add different additional enhanced features in the year 1998 and 2000. “H.264 is one of the popularly used codec on the planet, with significant note in optical disc, broadcast process, and streaming in video markets etc.[5] The applications are noted in Table I. Still, many uses of H.264 are subject to royalties, something that should is taken into considered before Google's WebM, as well as the general availability of decoding abilities on target platforms and devices” [8]. H. 264 mostly called as AVC (Advanced video coding), its block segmentation based, motion compensated with DCT technique. The aim behind AVC was to transfer video in low bit rate with better efficiency for UHD videos to its adaption of it.

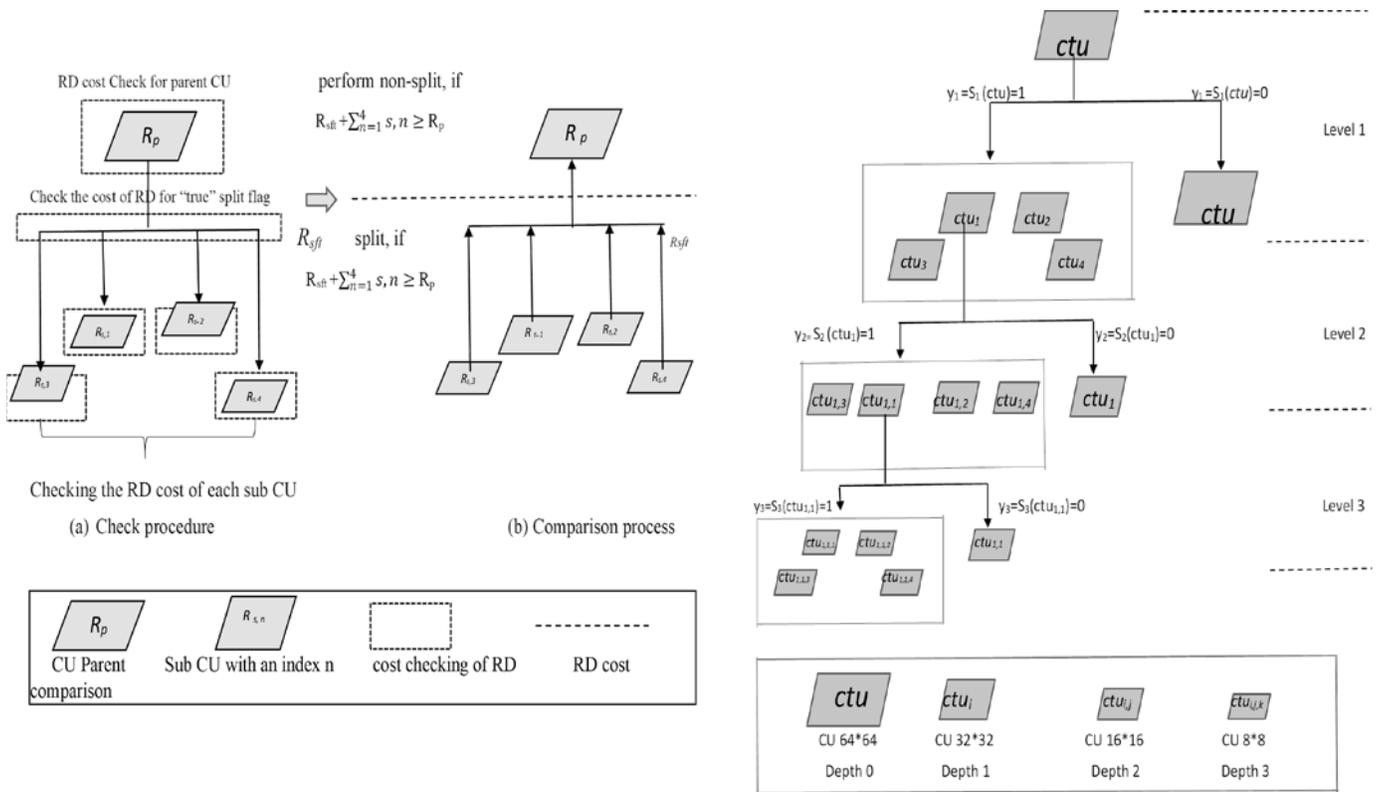


Fig. 1. (a) Rate Distortion Cost Calculation of CU Procedure. (b) The Representation of CU Classification as Layers to Analyze Depth.

“High Efficiency Video Coding, is also known as HEVC or H.265, is the step in this evolution. It builds off a lot of the techniques used in AVC/H.264 to make video compression even more efficient. When AVC looks at multiple frames for change those macroblock chunks can be a few different shapes and sizes, up to a maximum of 16 pixels by 16 pixels. With HEVC, those chunks can be up to 64x64 in size much larger than 16x16, which means the algorithm can remember fewer chunks, thus decreasing the size of the overall video” [9]. HEVC’s quad tree [10] partitioning uses the brute force search for RDO (rate distortion optimisation) cost calculation. The complexity of the procedure is more when used with normal signal processing steps that makes the HEVC [11] complex. Fig. 1(a) shows the procedure of rate distortion optimisation as a flowchart. It is divided into check procedure and comparison procedure. It initially checks for the rate distortion cost of the parent CTU [12] and the total cost of splitting it till end. Once this procedure is done comparison is done. In comparison it will the checking the RD [11] cost of parent and the cost after splitting. if the RD cost after split is more than the system will not split further and if the RD cost of parent is more then it will proceed with the split. This calculation procedure in HEVC is tedious that make the system complex. This issue was addressed by many algorithms, some provides

enhancement to the existing HEVC system while other set provides a totally new algorithm [3] providing a new architecture [13] to perform the procedure of compression. Deep learning based algorithms [13] [14] started working on this in recent years. So a depth decision algorithm with deep CNN [15][16] is modelled to solve this issue. Fig. 1(b) shows the level and depth of CTU. Understanding this depth concept [6] helps in designing deep CNN [13] algorithm to predict depth and thus to make intra prediction less complex.

The paper aims in complexity reduction in video compression (HEVC) by reducing encoding time. It is achieved by designing a deep learning-based system that predicts the depth of the CTU by making the intraprediction procedure less complex. The design is evaluated by pipelining it with the original HEVC and evaluates the complexity of the system. The overall design idea is shown in Fig. 2. The paper is divided mainly into two halves, 1) design of the deep depth decision algorithm, here the deep depth decision algorithm is designed tested and validated for datasets and 2) evaluation and experimental results of the model pipelined with original HEVC, were the model is pipelined with the original HEVC and the performance is evaluated for various resolution videos. The paper is concluded with the results showing encoding time reduction and future scopes.

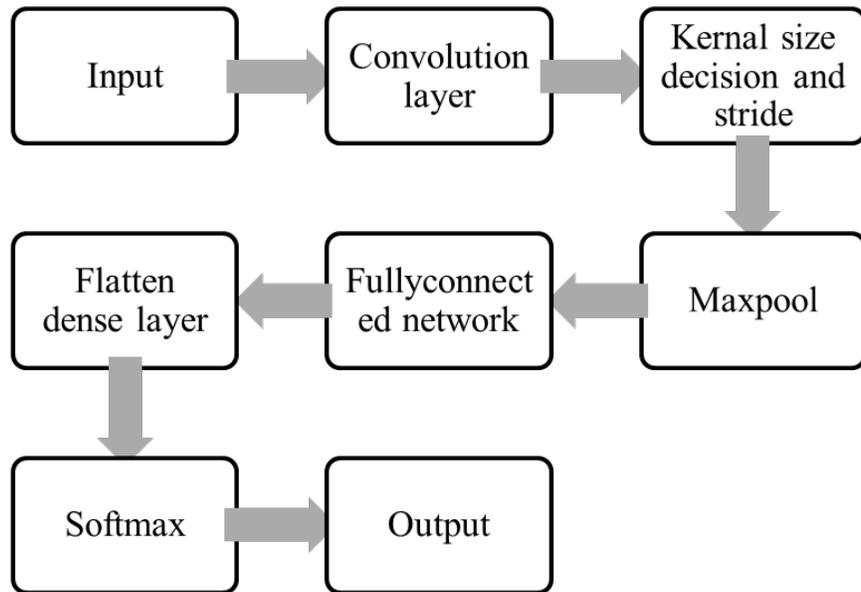


Fig. 2. Illustration of Steps in Order for the Modelling of Deep Depth Decision Algorithm with CNN.

## II. DESIGN OF DEEP CNN DEPTH DECISION ALGORITHM FOR INTER PREDICTION

The inter prediction and its computational complexity was the issue took for analysis to model a new network. The design of this network should possess less computational complexity compared to the existing system and should be

compatible to the existing codec. The design chosen should be compatible for faster transmission of frames while coding, so scalability and compatibility will also be the focus while designing, considering all this convolutional neural network (CNN) is chosen for this purpose so that all features are extracted correctly from the frames to produce better prediction as shown in Fig. 3.

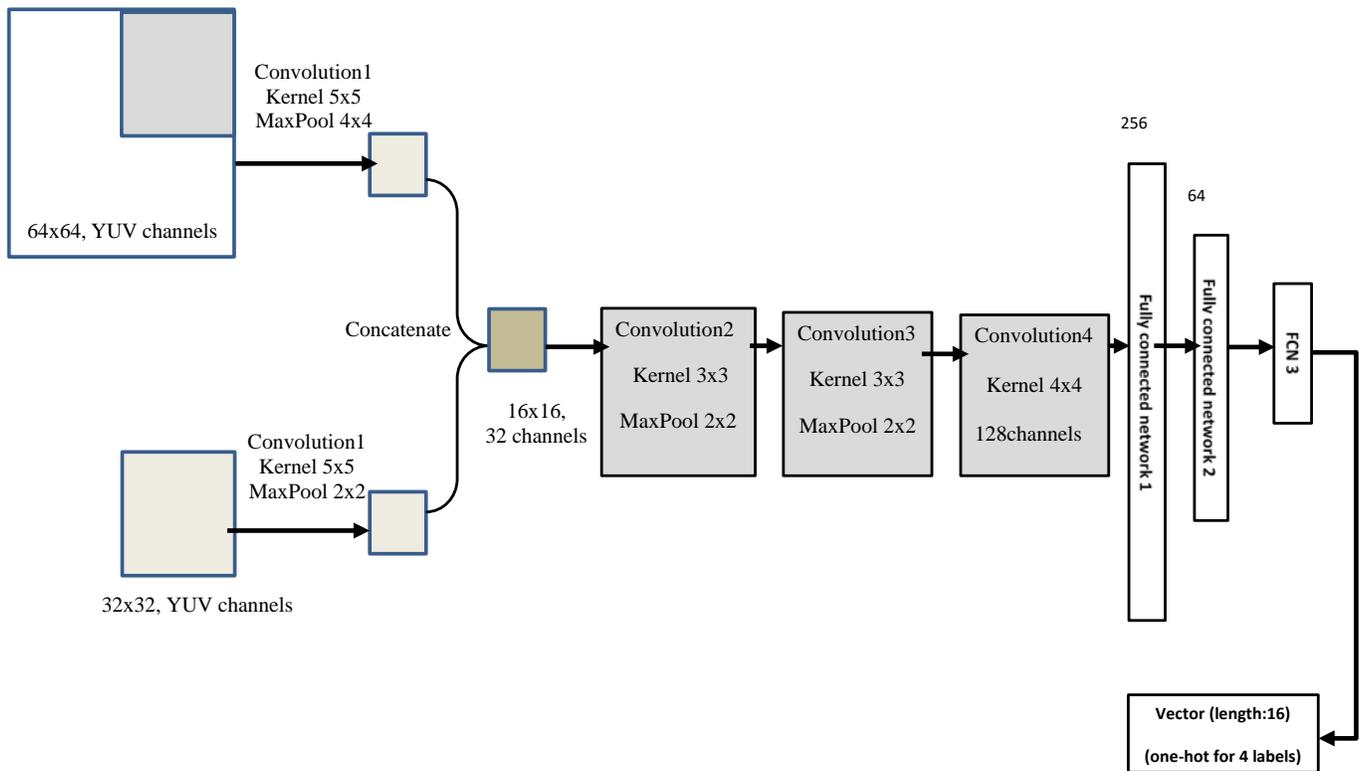


Fig. 3. The Representation of Deep Depth Decision Algorithm Model, with Input as 64X64 Patch followed by Convolution, MaxPool, Fully Connected Network followed by the 16 Length Output.

The CNN[17] used here is having multiple layer, the initial layer is the input layer. The input used here is video frames. The video frame can be of various properties, the YUV is the format chosen for this evaluation, other formats are also compatible in this model. The next layer in CNN model is convolutional layer, this is the layer that extracts the features of the frame based on the kernel used. The kernel size can be chosen based on the features that need to be extracted, if the kernel size is big it collects the global features or information from the frame whereas the small kernel [12] extracts the local features. Based on the need of the feature kernel can be chosen. In the design 5X5, 3X3 along with the 16X16, 4X4 [18] are also used, so model clearly extracts the global and local features from the frame. To cover all the inputs zero padding is used in this model. The stride used is same as the width of the kernel used in each case in the design. After extracting the feature its max pooled to reduce the size and converge the multiple values to a single value or less values. The activation function helps to decide the neuron is fired or not, so activation function is the node is kept in between and in the end of neural network. Here the activation function [19] used is ReLU rectified linear unit. ReLU maintains a value between (0-  $\infty$ ) zero and infinity by avoiding negative values. It's a simple function that returns if input is negative else returns the same value in other cases. Both forward and backward propagation exist in CNN [20] network. Here in the model training uses backward propagation while validation uses forward propagation.

The model designed takes the input as YUV CTU of 64X 64, the first convolution layer users its kernel and converts as 32X32 coding unit. The both CU of 32x32 are concatenated to extract more features and its pooled to 16X16 patch. The next stage of convolution with 3x3 kernel extracts its fine features, and a 4x4 for global features. After feature extraction in each stage the data are pooled by 2x2. In final stage the fully connected later flatten the information and compress it using SoftMax to 256 to 64 to a 16-length vector holding all the information of the CTU depth. The model is trained with various resolution input varying from 240p to 4k. After training the model will be having a training loss factor of 3.1049. the loss function estimated in this model is the cross entropy. Cross-Entropy Loss can be evaluated for separate images and independently and finally added together to obtain the final cross entropy as each path are mutually independent. A 66.12% of accuracy is obtained by the trained model.

#### A. Dataset

The dataset is the collection of sample video frames used for testing training and validation of the design proposed. Multiple and verity in dataset helps in the improvement of accuracy in the model. Here the dataset contains the Coding Unit image file extracted from the YUV video files as set of input and their corresponding depths for HEVC intra-prediction as output to train the proposed system. The dataset chosen here has multiple resolution and are not of same pattern videos to maintain the quality and efficiency of the model.

In HEVC intra-prediction, each I-frame is divided into 64x64 (CTU). For each 64x64 CTU, there's a depth prediction

represented by a 16x16 matrix. The elements in the matrix are 0, 1, 2 or 3, indicating depth 0/1/2/3 for a 4x4 block in the CT. The dataset contains images and corresponding labels. There're three folders: train, validation, test Image files: Each image may have different size based on the resolution of the video, and it is one frame extracted from a video. While using in the system, split the image into several 64x64 images or 32x32 and so on.

Labels: The labels are in separate folder called pkl folder. For one CTU, which is a 64x64 image file, the label will be a Python list with a length of 16. The length is 16 vectors instead of a 16x16 matrix, because there's redundant information for a 16x16 matrix, and it can be reduced to a 16x1 vector. So, for a 64x64 CTU, it has 16 labels, each label corresponds to a 16x16 image block in the CTU. If the frame is split into 64x64 CTUs, the size of the train dataset is around 110K images. The size of the validation dataset is around 40K images. The name of the image file will be like: v\_0\_42\_104\_.jpg, where v represents Video Number, followed by FrameNumber, CTU number and image extension. The Video Number is to find the corresponding .pkl file, like v\_0.pkl. To get the label for a certain 64x64 CTU, index the dict by:

label\_vector = video\_dict [FrameNumber][CtuNumber],  
for example: label\_vector = video\_dict ["42"] ["104"]. The label\_vector will be a length 16 Python list. Dataset loading in deep learning projects implemented in PyTorch can be done by in load\_example.py.

In the final stage for evaluation and comparison CPIH data set is also used to know the performance of the proposed system verses the existing models. The CPIH data set is not used in any of the testing or training for proper quality check.

#### B. Input and Pre-Processing Layer

The input used here is the YUV image patch derived from video frames. Each of this will be saved in a folder with separate labels in a python dictionary. The raw inputs need to be pre-processed by down sampling and splitting into 64x64, 32x32 and so on.

#### C. Convolution Layer

This layer performs convolution operation between the input and the kernel. If  $i$  is the pre-processed input and  $k$  is the kernel of size varying from 5X5 ,4x4,3x3 etc the convolution block output can be formulated as equation 1 and  $*$  represent the convolution operation. The size of the kernel decides the nature of the feature extracted.

In the design both global and fine features are extracted with variant kernels.

$$output = i * k \quad (1)$$

#### D. Fully Connected Layer

The fully connected layer initially flattens the output of convolution layer to a large single dimensional vector. The SoftMax operation helps to compress it further to required size without losing the information in it. The FCN1, FCN2 and FC3 along with averaging help it to shrink to 16 length vectors changing from 256 to 64 to 16.

E. Other Layers

The system is having a loss of feature dropping as the stages are crossing so the activation function ReLU[21][22], rectified linear unit shown in equation 2 where z is the input and R(z) is the output of ReLu . It should be noted that the output is activated by sigmoid function represented by S(z) in equation 3.

$$R(z) = \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases} \quad (2)$$

$$S(z) = \frac{1}{1+e^{-z}} \quad (3)$$

In original HEVC the prediction process is complex and time consuming as it should predict the RDO cost, so here the CNN network [23],[24]with depth decision [25][26] helps to predict the depth of each patch of 64X64 to a 16-length vector whereas original HEVC needs a matrix of size 64X64 to store it. The model converts the input patch to a 16 vector which can predict all the characteristics of that CTU with depth information as 0/1/2/3. The model is designed to take the input as 64x64 but while processing its split into 32x 32. Predicting for 64x64 patch directly doesn't make sense so the actual input is 32x32. The depth is 0 when the patch is not split and encoded as it is. The 64x 64 patch represent 16 length vectors.

So, for representing 32x32 the vector required is 4x4. So, in the output blocks of four 4x4 patches will be available as output for a CTU of 64x64. Each value in the vector indicate the depth of the CTU. if the first vector is 0 it says that It is a 64x64 patch and if depth is 1 means the 64x64 CTU is split once into four 32x32 CTUs and so on.

III. EVALUATION AND EXPERIMENTAL RESULT ANALYSIS OF DEEP CNN DEPTH DECISION ALGORITHM FOR INTER PREDICTION

The designed model is allowed to work with the HEVC codec as shown in Fig. 4. To simulate the original HEVC, HM software is used. The evaluation is done between the original HEVC and proposed model for intraprediction, pipelined to HEVC using CPIH dataset. Integrating neural network models in HEVC encoder helps to test the complexity reduction using deep-learning-based method in HEVC intraprediction. Using neural networks, the system can directly predict the Coding Unit (CU) depths for each frame. The intention is to speed up the encoding process of HEVC encoder. Thus, after we have a trained model, another thing that needs to be done is to integrate the deep learning prediction process into the HEVC encoder.

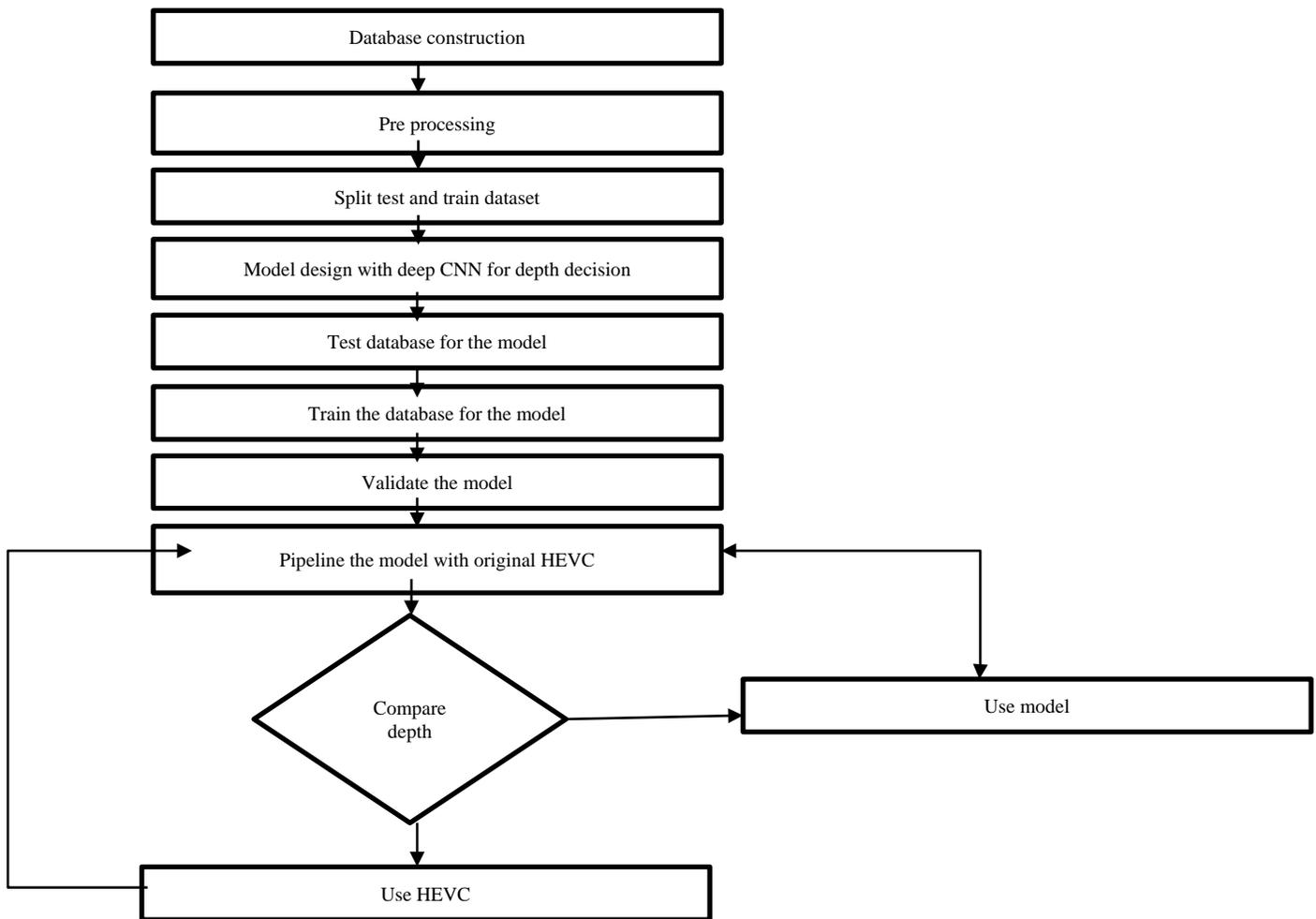


Fig. 4. Pipelining Structure with Deep Depth Decision Algorithm Model Added to the Original HEVC Model. It shows the Overall Steps for the Evaluation of the Designed Model.

This is to check the compatibility of the model with existing HEVC and also it makes the evaluation of the performance of our neural network model easier. This pipeline is used for evaluating the performance of a neural network model in HEVC intra-prediction process. Comparing the difference in encoding time, Y-PSNR, U-PSNR, V-PSNR, YUV-PSNR with the original HEVC encoder helps to know the efficiency of the model. FFmpeg, Python3, PyTorch are the requirements to perform this.

The frames are send for encoding to HM software for original HEVC encoding and calculates encoding time, Y-PSNR, U-PSNR, V-PSNR, YUV-PSNR and the same set is send to the pipelined model or the proposed model and calculates the encoding time, Y-PSNR, U-PSNR, V-PSNR, YUV-PSNR. Both are evaluated and made as graphical representation to check the performance comparison.

The results show that the time of encoding with and without pipelining the deep CNN network is shown in the Table I for some sample input. The input chosen for the test is mutually independent from the training set to maintain the accuracy and the wide range of resolution is also considered to check the performance of the system foe different resolution video frames. The results clearly show that there is a change in encoding time and thus the system proves it can reduce the and bit rate in each case, it supports the encoder with a better performance. Computational complexity of the original HEVC is high due to the RDO cost calculation. The experimental results show the time of encoding is drastically changed to low values for the proposed method. The PSNR curve is slightly

low here compared to original model but the system performance is not affected by this. The total process is done in python environment, when it's done, it will output with all information on the command line, like the encoding time, YUV-PSNR and so on. A sample output is shown in Fig. 5 and the comparison graphs are shown in Fig. 6. The proof of reduction in complexity is shown in Fig. 7 with the change in encoding time.

TABLE I. ENCODING TIME COMPARISON

| Image source                     | Resolution | Time of encoding with HEVC pipelined with Deep CNN network(T1) | Time of encoding with original HEVC(T2) | T1-T2     | $\Delta T$ proposed |
|----------------------------------|------------|----------------------------------------------------------------|-----------------------------------------|-----------|---------------------|
| CPIH dataset                     | 768x512    | 71.273                                                         | 664.634                                 | -593.361  | -89.27              |
|                                  | 1536x1024  | 467.671                                                        | 2741.481                                | -2273.81  | -82.94              |
|                                  | 2880x1920  | 2741.481                                                       | 16417.71                                | -13676.2  | -83.30              |
|                                  | 4928x3264  | 1910.618                                                       | 122731.3                                | -120820.7 | -98.44              |
| Average % $\Delta T$ improvement |            |                                                                |                                         |           | 88.49               |

```

CPIH007
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 719.1248 | 36.8114 | 37.8869 | 39.7502 | 37.4827

Y-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 305.1248 | 36.8114 | 37.8869 | 39.7502 | 37.4827

U-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

V-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

QPI: 0.000
Bytes written to file: CPIH007 (3059.536 Kbps)
Total Time: 71.273 sec.

```

a

```

CPIH007
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 11071.3688 | 37.7057 | 39.4391 | 40.5626 | 39.8866

Y-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 11071.3688 | 37.7057 | 39.4391 | 40.5626 | 39.8866

U-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

V-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

QPI: 0.000
Bytes written to file: CPIH007 (11071.369 Kbps)
Total Time: 466.671 sec.

```

b

```

CPIH007
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 12607.5288 | 37.8869 | 40.8209 | 41.8939 | 40.8428

Y-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 12607.5288 | 37.8869 | 40.8209 | 41.8939 | 40.8428

U-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

V-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

QPI: 0.000
Bytes written to file: CPIH007 (12607.529 Kbps)
Total Time: 2667.671 sec.

```

c

```

CPIH007
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 52894.4368 | 37.8708 | 42.0951 | 43.8317 | 40.8489

Y-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
20 | 52894.4368 | 37.8708 | 42.0951 | 43.8317 | 40.8489

U-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

V-Slices
Total Frames | Bitrate | Y-PSNR | U-PSNR | V-PSNR | YUV-PSNR
0 | nan(nan) | nan(nan) | nan(nan) | nan(nan) | nan(nan)

QPI: 0.000
Bytes written to file: CPIH007 (52894.437 Kbps)
Total Time: 122731.305 sec.

```

d

Fig. 5. Output Window showing the Bitrate-PSNR, U-PSNR, V-PSNR, YUV-PSNR, of Video Frame for Resolution (a) 768x512, (b) 1536x1024, (c) 2880x1920,(d) 4928x3264 with the State-of-Art Method and by HEVC-HM Software Simulation.

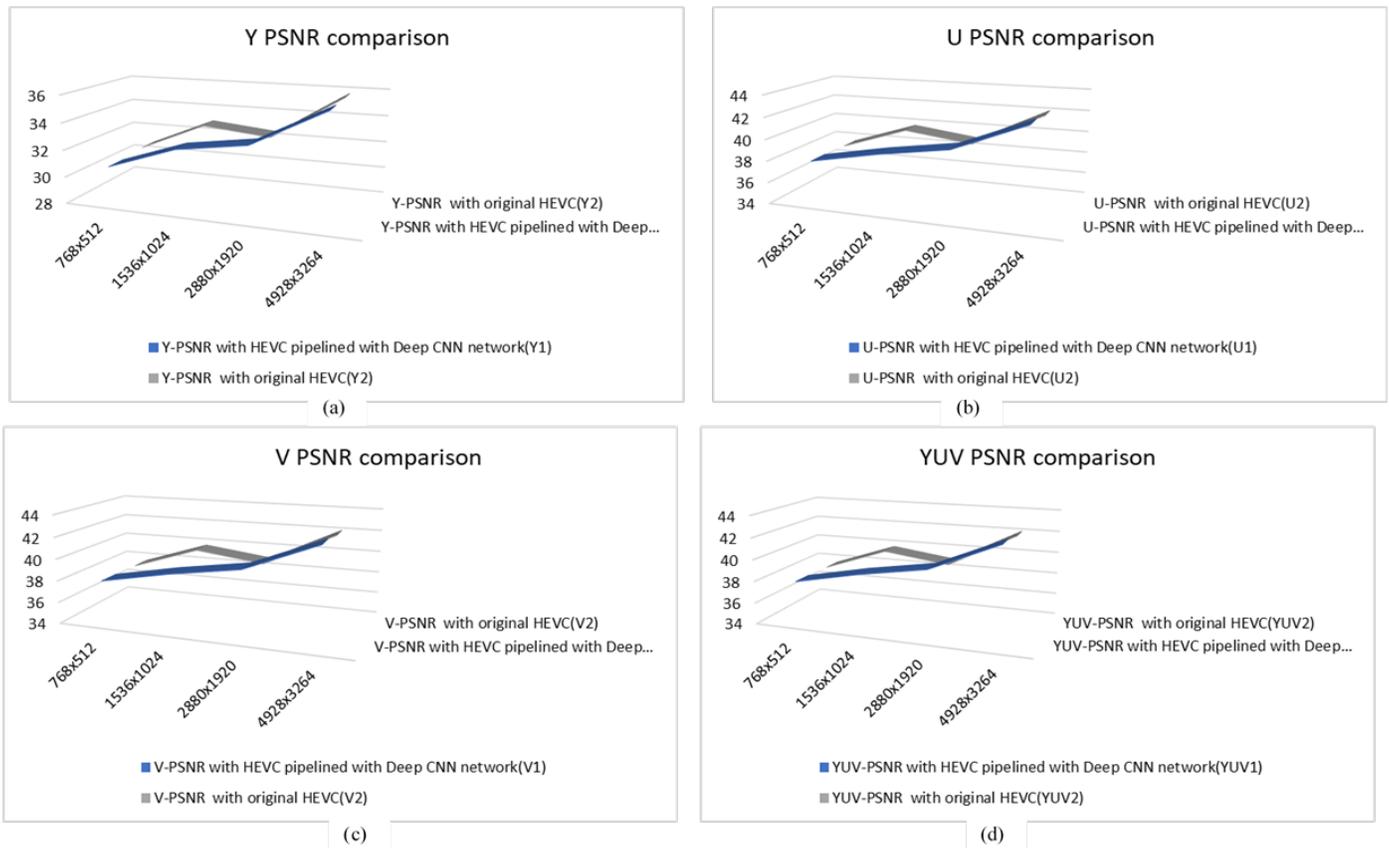


Fig. 6. Comparison Chart Showing (a) Y-PSNR, (b) U-PSNR (c) V-PSNR (d) YUV-PSNR, with the State-of-Art Method and by HEVC-HM Software Simulation for Video Frames of Different Resolution.

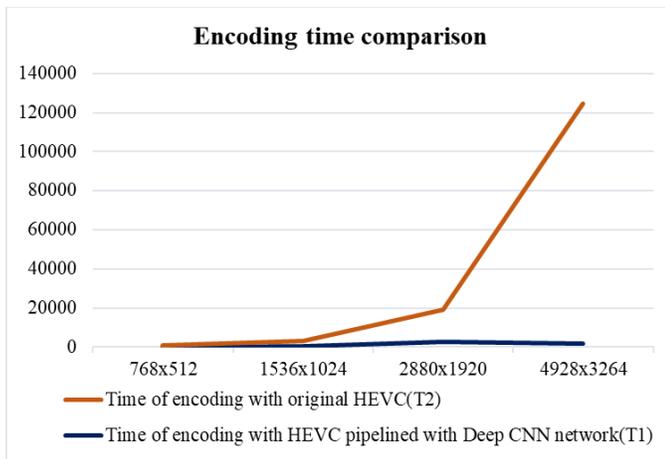


Fig. 7. Encoding Time Comparison.

#### IV. CONCLUSION

In this paper, a deep learning based inter-prediction is proposed to avoid the computational complexity issue in HEVC which predict the depth of the CTU in a 16-length vector than calculating the RDO cost by traditional signal processing method. The modelling adopted CNN network to perform this model with deep layers to predict the depth. The data set used for training was YUV and its tested on CPIH dataset to maintain the accuracy of the system and to avoid transfer or copied learning. The trained model is converted to

system and pipelined to the original HEVC system to check the performance. The system evaluated the time of encoding with and without pipelining and calculated  $\Delta T$ . The results and simulation clearly show that the design suits for the HEVC to work with less encoding time thus by reducing the complexity of the HEVC. The results prove it, the future enhancement on this can focus on the extension of this to inter prediction that improve the HEVC more.

#### REFERENCES

- [1] D. A. and N. G., "Combined spatial temporal based In-loop filter for scalable extension of HEVC," *ICT Express*, vol. 6, no. 4, pp. 306–311, 2020.
- [2] H. K. Joy and M. R. Kounte, "An Overview of Traditional and Recent Trends in Video Processing," *2019 International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2019, pp. 848–851, doi: 10.1109/ICSSIT46314.2019.8987896.
- [3] X. Li and N. Gong, "Run-Time Deep Learning Enhanced Fast Coding Unit Decision for High Efficiency Video Coding," *J. Circuits, Syst. Comput.*, vol. 29, no. 3, pp. 1–19, 2020.
- [4] G. J. Sullivan, S. Member, and T. Wiegand, "Video Compression—From Concepts to the H.264 AVC Standard.pdf," vol. 93, no. 1, 2005.
- [5] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, 2012. Joy.
- [6] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wanga, "Image and Video Compression with Neural Networks: A Review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8215, no. SEPTEMBER 2018, pp. 1–1, 2019.

- [7] J. L. Lin, Y. W. Chen, Y. W. Huang, and S. M. Lei, "Motion vector coding in the HEVC Standard," *IEEE J. Sel. Top. Signal Process.*, vol. 7, no. 6, pp. 957–968, 2013.
- [8] Eirina Boursoulatzé, Aaron Chadha, Ilya Fadeev, Vasileios Giotsas, and Yiannis Andreopoulos. "Deep Video Precoding", *IEEE Trans. Circ. and Sys. for Video Technol.* 30, 12 (Dec. 2020), 4913–4928. DOI:<https://doi.org/10.1109/TCSVT.2019.2960084>.
- [9] D. Liu, Z. Chen, S. Liu, and F. Wu, "Deep Learning-Based Technology in Responses to the Joint Call for Proposals on Video Compression with Capability beyond HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 5, pp. 1267–1280, 2020.
- [10] J. Lainema, F. Bossen, W. J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1792–1801, 2012.
- [11] S. Bouaafia, R. Khemiri, F. E. Sayadi, and M. Atri, "Fast CU partition-based machine learning approach for reducing HEVC complexity," *J. Real-Time Image Process.*, vol. 17, no. 1, pp. 185–196, 2020.
- [12] C. Ma, D. Liu, X. Peng, L. Li, and F. Wu, "Convolutional Neural Network-Based Arithmetic Coding for HEVC Intra-Predicted Residues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1901–1916, 2020.
- [13] J. K. Lee, N. Kim, S. Cho, and J. W. Kang, "Deep video prediction network-based inter-frame coding in HEVC," *IEEE Access*, vol. 8, pp. 95906–95917, 2020.
- [14] B. S. Kumar and V. U. Shree, "An End-To-End Video Compression Using Deep Neural Network," *JAC : A Journal of Composition Theory* ISSN : 0731-6755 vol. XIII, no. Xi, pp. 209–215, 2020.
- [15] O. Alharbi, "A Deep Learning Approach Combining CNN and Bi-LSTM with SVM Classifier for Arabic Sentiment Analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 6, pp. 165–172, 2021.
- [16] P. R. Lai and J. S. Wang, "Multi-stage Attention Convolutional Neural Networks for HEVC In-Loop Filtering," *Proc. - 2020 IEEE Int. Conf. Artif. Intell. Circuits Syst. AICAS 2020*, pp. 173–177, 2020.
- [17] S. Katiyar and S. K. Borgohain, "Comparative evaluation of CNN architectures for image caption generation," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 793–801, 2020.
- [18] L. Zhao et al., "Enhanced CU-Level Inter Prediction With Deep Frame Rate Up-Conversion For High Efficiency Video Coding" Institute of Digital Media & Cooperative Medianet Innovation Center , Peking University , Beijing , China Department of Computer Science , City Univiers," 2018 25th IEEE Int. Conf. Image Process., no. Mv, pp. 206–210, 2018.
- [19] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: a review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, no. 1, pp. 1–27, 2019.
- [20] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning Convolutional Networks for Content-Weighted Image Compression," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 3214–3223, 2018.
- [21] H. K. Joy, M. R. Kounte and A. K. Joy, "Deep Learning Approach in Intra -Prediction of High Efficiency Video Coding," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), 2020, pp. 134-138, doi: 10.1109/ICSTCEE49637.2020.9277189.
- [22] Joy H.K., Kounte M.R. (2022) Deep CNN Depth Decision in Intra Prediction. In: Subramani C., Vijayakumar K., Dakyo B., Dash S.S. (eds) Proceedings of International Conference on Power Electronics and Renewable Energy Systems. Lecture Notes in Electrical Engineering, vol 795. Springer, Singapore. [https://doi.org/10.1007/978-981-16-4943-1\\_1](https://doi.org/10.1007/978-981-16-4943-1_1).
- [23] Amitha I C, N S Sreekanth and N K Narayanan, "Collaborative Multi-Resolution MSER and Faster RCNN (MRMSER-FRCNN) Model for Improved Object Retrieval of Poor Resolution Images" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 12(12), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0121270>.
- [24] Naga Deepti Ponnaganti and Raju Anitha, "Feature Extraction based Breast Cancer Detection using WPSO with CNN" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 12(12), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0121250>.
- [25] Sigit Widiyanto, Dheo Prasetyo Nugroho, Ady Daryanto, Moh Yunus and Dini Tri Wardani, "Monitoring the Growth of Tomatoes in Real Time with Deep Learning-based Image Segmentation" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 12(12), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0121247>.
- [26] Shridevi Jeevan Kamble, Manjunath R Kounte, "SG-TSE: Segment-based Geographic Routing and Traffic Light Scheduling for EV Preemption based Negative Impact Reduction on Normal Traffic", *International Journal of Advanced Computer Science and Applications*, Vol. 12, No. 12, 2021, pp. 274-283.

# The Pragmatics of Function Words in Fiction: A Computer-aided Text Analysis

Ayman Farid Khafaga

College of Science and Humanities, Prince Sattam Bin Abdulaziz University, Saudi Arabia  
Faculty of Arts and Humanities, Suez Canal University, Egypt

**Abstract**—This paper uses a computer-aided text analysis (CATA) to decipher the ideologies pertaining to function words in fictional discourse represented by Edward Bond's *Lear*. In literary texts, function words, such as pronouns and modal verbs display a very high frequency of occurrence. Despite the fact that these linguistic units are often employed to channel a mere grammatical function pertaining to their semantic nature, they, sometimes, exceed their grammatical and semantic functionality towards further ideological and pragmatic purposes, such as persuasion and manipulation. This study investigates the extent to which function words, linguistically manifested in two personal pronouns (I, we) and two modal verbs (will, must) are utilized in Bond's *Lear* to convey both persuasive and/or manipulative ideologies. This paper sets three main objectives: (i) to explore the persuasive and/or manipulative ideologies the four function words under investigation communicate in the selected text, (ii) to highlight the extent to which CATA software helps in deciphering the ideological weight of function words in Bond's *Lear*, and (iii) to clarify the integrative relationship between discourse studies and computer-aided text analysis. Two findings are reported in this paper: first, function words do not only carry semantic functions, but also go beyond their semantic functionality towards pragmatic purposes that serve to achieve specific ideologies in discourse. Second, the application of CATA software proves useful in extracting ideologies from language and helps better understand the power of function words, which, in turn, accentuates the analytical integration between discourse studies and computer, particularly in the linguistic analysis of large data texts.

**Keywords**—Computer-aided text analysis (CATA); concordance; function words; persuasion; manipulation; ideology; Bond's *Lear*

## I. INTRODUCTION

Starting from the assumption that language is a means of communication that often reflects the ideologies of its users, it can be claimed that there is a reciprocal relationship between language and ideology [1]. This relationship has been approached within the field of fictional discourse [2], and in other discourse genres [3]. The different linguistic units expressing language can also be said to be ideology carriers. In the realm of critical discourse analysis (CDA), ideology is one cornerstone of its analytical umbrella to the extent that it is uncommon to conduct a CDA for any text without a reference to ideology; CDA takes as one of its core concerns the task of decoding the hidden ideologies in discourse, either spoken or written. According to van Dijk [4], ideology refers to such set of specific rules, beliefs, and attitudes that are commonly shared between individuals of the same group,

institution and/or party. For him, ideologies are individualized - and institutionalized-based notions that not only demarcate the process of communication within the in-group, but also determine specific communicative guidelines to contact with the out-group. These shared-beliefs often serve to discursively distinguish between the 'We' and 'They' relationship in discourse. The communal relationship between language and ideology is clearly evident in discourse, particularly where notions of power, dominance, and control are addressed [5]. Ideology can be discussed from different linguistic angles, including the semantic, the pragmatic, the lexical, or the grammatical, and it can also be communicated at the different levels of discourse: the word, the sentence, and the utterance. Even function words, such as pronouns and modal verbs, whose main purpose in discourse is to convey a grammatical function, can also be perceived as ideology carriers.

This paper attempts to decipher the ideological significance of function words in Edward Bond's *Lear*. The reason why a drama text is selected for the analysis lies in the fact that literary genres always witness numerous number of function words that are recurrently repeated in these texts. In many cases of their usage in texts, these words carry their naturally semantic function of grammaticality. So, for instance, the different personal pronouns are employed to conduct their semantic function of just referring to deictic concepts, and also the modal verbs can be discursively used to indicate obligation (must, should), high level of certitude (will), possibility (can), etc. In light of this paper, these function words are linguistically investigated by means of a computer-aided text analysis (CATA) and CDA to decode the ideologies these words convey beyond their ordinary semantic functions.

As a result of the incessant technological development, computer software have come to occupy substantial significance in numerous studies within the scope of linguistic studies, as they are used to draw both theoretical and empirical results that contribute to the field of linguistics in general and to textual analysis in particular [6], [7], [8], [9]. These studies highlighted the indispensable role of computer software as digital tools that serve to support and facilitate a comprehensive and enhanced analytical milieu, wherein analysts and linguists can easily manage their analyses by providing adequate, credible and ample results. According to [10], adopting a computational approach to the analysis of fictional texts not only facilitates the whole process of text analysis, but also emphasizes the integration between modern technologies and other social and human disciplines.

The paper is theoretically framed upon three concepts. The first is critical discourse analysis, which is approached in terms of Fairclough's [11] model of analyzing grammatical concepts in discourse, including pronouns and modality, the core concern of this study; the second is ideology, which is addressed in light of van Dijk's [12] perspective concerning the concept; and the third is a computer-aided text analysis, which is represented by a frequency distribution analysis (FDA) to the four function words under investigation. Crucially, the analysis seeks to highlight the analytical integration between CDA and CATA, as well as to shed light on the way these two analytical tools are interwoven within the discourse of Bond's *Lear* to decipher the hidden ideologies of persuasion and/or manipulation encoded in the conversational turns of characters in the selected play, and channeled by the function words employed throughout dialogicity.

Approaching pronouns and modality as carriers of ideology in discourse by the application of CATA via concordance reflects the significance of the grammatical aspects and function words as linguistic tools in the communication process [13], and the significance of utilizing and applying computer software to the analysis of large data texts [14], [15]. There is no discourse that does not carry ideological significance; such an ideological significance serves to open the gate of research towards recurrent discussions that function to discover further meanings pertaining to texts, or to challenge and refute the existing meanings of such texts. Literature is a fertile soil wherein discourse analysts and linguists find so many linguistic phenomena worthy of linguistic research [16], [17]. Despite the fictional nature of communication in literary texts, they are still considered as mirrors of what is going on reality, and, therefore, are perceived as parallel to naturally occurring conversations.

#### A. Research Questions

The current study tries to offer answers to the following research questions:

- 1) To what extent does a computer-aided text analysis contribute to the analysis of fictional texts?
- 2) What are the ideologies communicated by function words in the selected play?
- 3) To what extent does Key Word in Context (KEWIC) offered by concordance contribute significantly to the understanding of the power of function words in fictional discourse?

#### B. Research Objectives

The answer of the abovementioned research questions constitutes the main objectives of the study as follows:

- 1) To highlight the extent to which computer-aided text analysis helps in deciphering the ideological weight of function words in Bond's *Lear*.
- 2) To clarify the complementary relation between critical discourse analysis and computer-aided text analysis.
- 3) To explore the persuasive and/or manipulative ideologies function words communicate in the text at hand.

In what follows, the paper provides the theoretical background as well as the review of literature relevant to the study of function words as carriers of ideologies in discourse in Section II. Section III provides the methodology adopted in this paper by offering the analytical procedures, the rationale, and the description of the selected data. Section IV is dedicated to the analysis of the selected data. Section V presents the discussion of the results reported in this study, and Section VI is the conclusion of the article, which also provides recommendations for further research.

## II. LITERATURE REVIEW

### A. Computer-Aided Text Analysis (CATA)

The employment of a computer-aided text analysis (hereafter, CATA) proves useful for the understanding of the thematic and ideological message of texts in corpus linguistics [18]. Applying the different computer software to the analysis of texts serves to facilitate the process of interpretation pertaining to these texts, which in turn, helps decode the ideological significance encoded in the linguistic expressions, either at the word level or the sentence one. Such ideological significance carried by words and/or sentences is difficult to be deciphered if it is approached manually; that is, without the work of computer [19]. Nowadays, computer software occupies an integral part in the field of linguistics. The importance of computer software is not only confined to the computational linguistics studies, but they have their contributive part in the other fields of linguistics, including pragmatics, semantics, morphology, and discourse studies [20]. This is because computer software such as concordance can efficiently foster the analytical process in large data texts in a way human performance alone proves to be inadequate [21].

It is worth mentioning that CATA offers various analytical tools and options. One of these analytical options is the Frequency Distribution Analysis (FDA), which entirely functions to provide the number of occurrences a searched item occurs in a text. According to [22], the frequency analysis that can be generated by concordance makes it available for analysts to have a general look about the textual nature of a specific lexis in a text. This further serves to direct the analytical wheel towards the significant precedence of one occurrence over another, which is computationally enabled by the second variable offered by CATA, that is, the variable of Key Word in Context (KWIC). The variable of KWIC provides the contextual picture in which a specific searched word occurs. In other words, KWIC clarifies the contextual environment of the searched items, which, in turn, helps arrive at the ideological significance pertaining to words and/or phrases [23]. A further analytical option realized by CATA is Content Analysis (CA). For Weber [24], content analysis serves to categorize the different words into classes according to their semantic features. This content or semantic categorization is very contributive to the thematic intelligibility of texts, as it classifies words into semantic groups that ultimately function to achieve a comprehensive understanding of the thematization of texts, particularly large data texts such as the literary ones, as is the case with the play under investigation.

In light of this paper, CATA is enabled by the program of Concordance to provide the analytical options listed above. Concordance is a computer software that enables analysts and users to collect, access, classify and analyze the different types of texts, specifically those that abound in large amount of data [25], [26]. Concordance, therefore, can retrieve all occurrences of a searched lexis in a text, can display the contextual environment of any word, and can categorize all words according to their semantic content [27]. By deriving the frequency distribution of the four function words under investigation, which is accompanied by both the use of KWIC and content analysis, the ideologies the four function words carry will be revealed.

### B. Ideology and Discourse

Ideology has been a main area of concern for discourse and ideological discourse analysts within linguistic studies adopting critical discourse analysis as their theoretical framework [28], [29], [30], [31]. These studies have emphasized the connection between language, power and ideology. They point out that ideology is enacted through language and helps to legitimate domination. They argue that any aspect of structure could be ideologically significant; that is, all linguistic expressions can communicate ideological significance, which can also be manifested phonologically, syntactically, semantically, pragmatically, etc. Highlighting the significant role of institutionalized ideology, van Dijk [32] postulates that ideologies are specific types of ideas that form what he terms 'belief systems' or 'social representations'. He maintains that such beliefs cannot only perceived as belonging to individuals, but they also form the general cognitive and background shared by groups.

According to van Dijk [33], ideology is a mediator element between society and discourse. Such a relationship serves to facilitate and activate the linguistic intelligibility among the different members of the group. Forming and agreeing upon a unified ideological background within the in-group is a prerequisite that helps to establish a successful act of communication produced by the different representations of discourse. By means of ideology, relations of power and dominance are motivated in discourse; power relations are produced and practiced. Ideology is employed also to persuade and/or manipulate; in both cases, it is a medium through which the powerful exercise their power and the powerless resist. Consequently, one can find different discursive practices that tend to manipulate and others to persuade. The targets sought beyond each type determine the type of ideology practiced in the process of communication, that is, whether or not the benefit is both speaker and hearer-oriented (persuasion), or only speaker-oriented (manipulation). Van Dijk maintains that ideology intelligibility among the members of the group leads to a successful participation of the individual member towards the ideological principles shared by the group as a whole. This ideological participation tends to create a discursive cohesion among discourse participants, which, for Khafaga [34], accentuates the idea that individually-based ideology affects and is affected by the institutionalized ones; it is an unremitting process of influence that emphasizes ideology-discourse reciprocity.

### C. Fairclough's Grammatical Approach to Critical Discourse Analysis

In discussing the role of function words and their ideological significance in discourse studies, Fairclough [11] proposes four sets of items that can be used for the linguistic analysis of function words in texts and discourse. The first set requires the investigation of the experiential values of texts, which necessitates the analysis of pronouns and modality. The second set involves the analysis of the relational values of texts, and entails a focus on the grammatical features, such as the type of sentence used, i.e. declarative, interrogative, or imperative; the type of modality: truth, obligation, or possibility modality; and the type of pronouns used in discourse. The third set of items constitutes the study of the expressive values grammatical features have, including expressive modality. The fourth set comprises the analysis of the different types of sentences used; for example, simple, compound, or complex, as well as the relationship between the various structures of sentences. Fairclough's sets of items abound in grammatical aspects relevant to produce an ideology-loaded type of discourse. Pronouns and modality are discursively employed to express, produce and maintain agency, particularly in the field of ideological/critical discourse studies [35], [36], [37]. These studies clarified that the reason lies in the fact that agency is closely related to notions of power and domination, and it is difficult to find any ideological discourse that does not address issues of power, dominance, persuasion, and manipulation.

### D. Pronouns

Using pronouns in discourse is perceived as one way of communicating agency, which, in turn, operates as conduits of ideologies in discourse [38], [39]. The use of the pronoun 'I' serves to communicate the competency, authority and responsibility of the speaker, whereas the pronoun 'you' conveys domination among discourse participants. Concerning the pronoun 'we', Fairclough [11] argues that it is divided into two types: "inclusive 'we' and exclusive 'we'". The inclusive 'we' includes both the reader and the writer (speaker and hearer), whereas the exclusive 'we' refers to the writer (speaker) only without any reference to the other participants in discourse. Pinto [40] further affirms that the use of the inclusive 'we' indicates that the goals and benefits of the whole group is more important than the benefits of the individual, whereas the exclusive 'we' establish a border between the benefits of the individuals and those of the group. Pinto maintains that in the case of manipulative discourse, the inclusive 'we', once employed, masks imposition under the guise of cooperation, as it manipulatively shows that the benefits of the group is inferior than those of the individual. Therefore, the inclusive 'we' is discursively used to show unity, solidarity and competency, whereas the exclusive 'we' indicates distance and separateness between the speaker and his/her addressees.

### E. Modality

According to [11], modality is classified into "relational modality" and "expressive modality." Such categorization, for him, is based on the type of authority exercised by the speaker over his/her addressees. In relational modality, the authority is

practiced by one participant over another, while in expressive modality; the focus is on the speaker's authority in terms of the truth of the propositional content of the linguistic expression. He maintains that it is not only through modal verbs that modality can linguistically be communicated, but it can also be expressed by other grammatical tools, including adverbs and tense. For Fowler [41], modality has four types: truth, obligation, permission and desirability. Truth modality can be expressed by modal auxiliaries represented in 'will', adverbs of certitude such as 'certainly'. This type of modality shows that the speaker's assumption is completely true. Truth modals are used to express a high level of certitude. Obligation modality can be realized by the use of some modal auxiliaries, such as 'must', 'should' or 'ought to'. Obligation modality draws the recipient's attention to the necessity of carrying out the speaker's proposition. Permission modality can be presented by the modal auxiliaries, such as 'can' and is used to give their participants permission to carry out their propositions. As for desirability modality, it concerns itself to clarify the speaker's status of accepting or refusing what is communicated by his/ her proposition.

#### F. Previous Studies

The use and application of computer software in general and concordance in particular to the analysis of large data texts such as the fictional ones has been the focus of many studies within the field of linguistics. [42], for example, used data mining to investigate the semantics, rhythm and pace in narrative writing. They clarified the extent to which input data can significantly contribute to the final interpretation of fictional texts, and concluded that data mining via visualization can mirror the semantic categorization a fictional text carry. Another study conducted by [43] explored the extent to which concordance is effective in the analysis of fictional discourse. This study demonstrated that concordance can be applied to large data texts in order to generate authentic and credible results that contribute to the interpretation of texts. This study focused on two analytical variables generated by concordance: frequency distribution analysis and key word in context. The study concluded that the application of concordance to the study of literary genres contribute to the thematic and ideological understanding of texts.

Within the scope of translation studies, [44] showed the significance of using modern technological software in producing credible versions of translations. The study recommended the incorporation of computer software into the process of teaching and learning university courses of translation in the different academic institutions in Saudi Arabia. Furthermore, [45] conducted a research in which they explored the impact of CALL software (computer assisted language learning) on the performance of students who major English as a foreign language (EFL) in Saudi university. This study clarified that the application of CALL to EFL settings influences the various learning outcomes of EFL students positively. The study is based on testing the effectiveness of using the two computer programs of Snagit™ and Screencast on acquiring the skill of reading. The study revealed that the application of the two computer software serves to improve the academic level of students, by fostering the linguistic skills pertinent to the acquisition of the skill of reading. The study

also reported that such technological incorporation into EFL course functions to develop not only the linguistic competence of EFL students, but also their communicative skills. This study concluded by recommending the application of the different CALL software to the different EFL courses in the context of Saudi universities.

Within courtroom settings and legal discourse studies, [46] conducted a research in which they explored the extent to which concordance helps investigate the linguistics of opening statements, by decoding the various ideologies beyond the semantic proposition of the linguistic expressions. This study employed a frequency distribution analysis to arrive at the hidden meanings and the pragmatic purposes targeted as a result of the recurrent employment of particular lexical items in the analyzed texts. The aforementioned studies show the effective and contributive role computer software play in the field of linguistic studies, either on fictional texts or outside the scope of fiction, i.e., in EFL and courtroom settings. Such contributive significance is anticipated to be extended in this study to present a further dimension of the application of CATA to decipher the ideologies of function words in fiction.

It is perspicuously evident that all previous studies have employed CATA software into the linguistic analysis of texts. Some of these studies focused on fictional texts, whereas other studies have presented discussions on legal texts, EFL settings, etc. One observation concerning related literature is that it did not use CATA software within the scope of pragmatics; that is, none of the previous studies has employed CATA software to explore the different pragmatic purposes in discourse. This last point is the core concern of the current study, which constitutes the research gape attempted to be addressed in this article.

### III. METHODOLOGY

#### A. Data and Rationale

The data in this study encompasses Edward Bond's *Lear* [47]. The play is structured around 4 acts that constitute eighteen scenes forming the whole production of the dramatic work. The rationale beyond the selection of this play in particular is due to the fact that it witnesses a significant usage of some grammatical aspects that prove indicative in communicating different ideologies, varying from the persuasive to the manipulative. This has been marked by the frequency analysis added in this article, which displays an ideological weight for such grammatical aspects; they are not employed chaotically in the dramatic dialogue of the play, but are vessels of ideology. Significantly, Bond's *Lear* exhibits a remarkable ideological discourse that requires a concise linguistic analysis, specifically with regard to the dexterous employment of pronouns and modality to communicate ideologies.

#### B. Research Procedures

Three procedural steps were followed in the analysis of the selected play. The three stages revolved around the use of three variables of CATA: frequency distribution analysis (FDA), key word in context (KWIC), and content analysis (CA). The first stage constituted preparing the text of the selected play, by uploading it electronically so as to be ready

for analysis. This stage offered a general idea of the way the discursive atmosphere of the play is communicated via the conversational turns of its characters. In the second stage the four function words (I, we, will, must) were electronically highlighted to mark their occurrence in the play. This was conducted by using a frequency distribution analysis to the whole play-text, wherein occurrences of each searched item were monitored. The third stage presented an interpretative task in which all highlighted items were discussed in terms of their indicative occurrences within the particular contextual environment in which they occur. After the three analytical stages, findings were firstly reported, and then interpreted in terms of the extent to which the searched items were contributive to conveying specific persuasive and/or manipulative ideologies. Significantly, all emphases (italicized) in the selected extracts in the analysis part are made by the author for analytical reasons.

C. The Frequency Analysis

During the process of analysis, the work of concordance was confined to reflecting a frequency distribution analysis for the searched lexical items that marked as indicative in the study of pronouns and modality as ideology carriers. This frequency analysis is conducted by concordance. Concordance facilitates the process of accessing and examining large data texts in order to arrive at credible and concise results that could be difficult to be realized if the analysis is conducted without the help of computer software [48], [49]. The options provided in concordance here were only to mark the word in its contextual environment. This context was determined by only five words before and other five words after the searched item. This functions to give a brief about the nature of the linguistic context in which the word occurred in text. Kennedy [50] argues that concordance is a software that serves to generate all occurrences of a given word or lexis in a corpus. Further, Hockey [51] points out that a concordance or a frequency analysis is produced by virtue of the searched item and the contextual environment in which it occurs. Concordance, for him, offers what is called KWIC (Key Word in Context), that gives much information about the searched word in its different contexts in text. This, in turn, serves to extend the interpretative process, opening new horizons that help better interpret the linguistic expression.

IV. DATA ANALYSIS AND RESULTS

This part constitutes two analytical strands: pronouns and modality, both of which reflect the extent to which pronouns and modality are carriers of specific ideologies, persuasive and/or manipulative in the discourse of the selected play.

A. Pronouns

This part of the analysis presents two pronouns that are used in the discourse of Bond's Lear to communicate specific ideologies varying from persuasion to manipulation: the pronoun 'I' and the pronoun 'we'.

1) *The pronoun I*: The pronoun 'I' is used in Bond's play to communicate specific ideologies, whose core concern varies from persuasive ideology to manipulative one. The former aims to get the addressees persuaded of some particular idea, whereas the latter constitutes the realization of

the speaker's desire even if it runs counter to the recipients' attitudes. Thus, one can say that the first type is addresser-/addressees-benefit oriented, while the second type is only an addresser-benefit oriented. Consider the following extract:

Lear. I knew it would come to this! I knew you were malicious! I built my wall against you as well as my enemies! (Lear (henceforth, L)., p. 7)

Lear manipulates the first-person pronoun 'I' to persuade his participants of the legitimacy of building his wall. He employs the deictic 'I' to show that he was right when he decided to build the wall. Lear's use of the past tense in *knew* and *built* also serves to prove his competency, and to clarify that he was far-sighted when he started to build the wall. Lear puts himself in the position of the agent and puts his two daughters in the position of the patient. Through using agency, Lear tries to communicate that it is he who has the competency to foresee the future and to give judgment. In this way he tries to make them accept everything he is going to allege. The following tables present a frequency analysis of the thematic use of the pronoun 'I', through which the different types of ideologies communicated by the first person singular 'I' can be caught by the context in which it occurs.

TABLE I. A FREQUENCY DISTRIBUTION ANALYSIS OF THE MANIPULATIVE 'I'

| I.....TF (555)            |      |                          |      |
|---------------------------|------|--------------------------|------|
| context                   | word | context                  | line |
| when I was young.         | I    | stopped my enemies       | 103  |
| my sworn enemies.         | I    | killed the fathers       | 176  |
| me. And when              | I    | killed the fathers I     | 178  |
| when I killed the fathers | I    | stood on the field among | 178  |
| This is not possible!     | I    | must be obeyed!.         | 238  |
| two bodies with them.     | I    | Knew it would come       | 289  |
| It would come to this     | I    | Knew you were malicious! | 289  |
| you were malicious!       | I    | Built my wall against    | 290  |
| That's how                | I    | crushed the fathers      | 341  |
| O why did                 | I    | cut his tongue out?      | 625  |

Note: TF means total frequency of the searched word

TABLE II. A FREQUENCY ANALYSIS OF THE PERSUASIVE 'I'

| I.....TF (555)      |      |                           |      |
|---------------------|------|---------------------------|------|
| Context             | word | context                   | line |
| So stay. LEAR .     | I    | could have a new life     | 1048 |
| wall down, and      | I    | had to stop that          | 2001 |
| hand.) And now      | I    | must move them            | 2003 |
| Murderer. And now   | I    | must begin again.         | 2518 |
| must begin again.   | I    | must walk through my life | 2519 |
| shivering in blood, | I    | must open my eyes         | 2522 |
| hurry on. LEAR.     | I    | am the King! I kneel      | 2765 |
| what she's doing!   | I    | must tell her             | 2798 |
| eyes, my eyes!      | I    | must stop her             | 2805 |
| all my mistakes,    | I    | understand all that       | 3255 |
| As fit as I Was.    | I    | can still make my mark    | 3662 |

As indicated from the two tables above, Table I shows that only 10 occurrences out of 555 are employed to convey manipulative ideologies. Table II clarifies that 11 occurrences are used to indicate persuasion. The contextual environment in which the deictic 'I' occurs reflects the extent to which it is employed to achieve persuasion and/or persuasion, which is clarified by the thematic and content analysis of the KWIC pertaining to the pronoun 'I'.

2) *The pronoun 'we'*: The pronouns 'we', 'us', and 'our' are used inclusively in Bond's play to establish relations of agreement, solidarity, and inclusion; and exclusively to show power, distinction and authority. These pronouns are among the main rhetorical strategies speakers used to communicate ideology, or as [52] puts it pronouns are "one of the major tools of persuasion used by politicians" (P. 37). Consider the following extract:

Cordelia. You were here when they killed my husband. I watched them kill him. I watched and I said we won't be at the mercy of brutes anymore, we'll live a new life you must stop speaking against us. (L., p. 83).

Cordelia tries to convince Lear to stop talking against the new government. She starts her persuasive ideology with clarifying a number of irrationalities committed against her on the hands of Bodice and Fontanelle. She involves Lear in her speech as a witness you were her when they killed my husband. Cordelia then uses the pronouns 'we' and 'us' both inclusively in we won't be at the mercy of brutes anymore and 'we'll live a new life; and, exclusively, in you must stop speaking against us to show intimacy and solidarity with the old king in the first two utterances, and to threaten the him in the third utterance so as to stop him talking against her. Cordelia's use of the first-person plural pronoun also indicates that she is authoritative enough to speak on behalf of others in the government, which reflects her power and domination. Cordelia's first utterance we won't be at the mercy of brutes anymore emphasizes her power and determination that she will never allow herself to be at the mercy of brutes again. Her second utterance we'll live a new life is an attempt to manipulate Lear's mind through a seductive promise of a new life in the future in which he will live in peace under her rule. The connection between the truth modal 'will' and the pronoun 'we' emphasizes her ability to carry out what she promises to do. Cordelia's exclusive 'us' in against us serves to show her power and domination over the situation. She establishes herself as a leader who has the complete access to speak on behalf of others in the government. Tables III and IV offer a frequency distribution analysis and a KWIC of both the inclusive and exclusive 'we'.

TABLE III. A FREQUENCY ANALYSIS OF THE INCLUSIVE 'WE'

| WE.....TF (104)    |      |                       |      |
|--------------------|------|-----------------------|------|
| Context            | Word | Context               | Line |
| How could          | we   | ever be free?         | 105  |
| and I said         | we   | Won't be at the mercy | 3478 |
| brutes any more,   | we   | will live a new life  | 3479 |
| much suffering But | we   | made the world        | 3531 |
| and fragile and    | we   | have only one thing   | 3532 |

TABLE IV. A FREQUENCY ANALYSIS OF THE MANIPULATIVE 'WE'

| WE.....TF (104)                |      |                               |      |
|--------------------------------|------|-------------------------------|------|
| Context                        | Word | Context                       | Line |
| CORDELIA. When                 | we   | have power these things won't | 1867 |
| don't build the wall. Cordelia | we   | must. Lear.                   | 3499 |
| Tell her . CARPENTER.          | We   | came to talk to you           | 3508 |
| There are things               | We   | have other opponents          | 3539 |
| and tell you this before       | we   | put you on trial              | 3542 |

The two tables show that the pronoun 'we' is used inclusively 5 times (Table III), and exclusively 5 times (Table IV) out of 104 occurrences. In both cases the pronoun 'we' is employed to carry manipulative ideology. Again, this manipulative usage of the pronoun 'we' is demonstrated through the variable of KWIC. That is, by looking at the contextual environment in which this pronoun occurs in text.

*B. Modality*

Two types of modality are discussed in this section: obligation modality and truth modality. Both types expresses agency and are carriers of ideology.

1) *Obligation modality*: The obligation modals 'must' and 'should' are dexterously employed in Bond's Lear to reflect the power of the speaker over his participants. Speakers use these modals to impose their own ideology over their recipients and to direct their behavior towards a complete obedience and submission to their own purposes [53]. The use of obligation modality dominates the discourse of oppression in which powerful characters use these modals to practice their domination over those who are powerless. Notice the following extract:

Bodice. We must go to our husbands tonight. We must attack before the wall's finished. We must help each other. (L., p. 8).

Bodice is talking to her sister, Fontanelle, with regard to their plan to attack Lear's army and to wrest him from his kingdom. Bodice uses the obligation modal 'must' three times to convey the necessity of doing what they decide to do, and to emphasize that attacking their father becomes urgent so as to stop the acts of building on the wall. Bodice's use of the obligation modals in we must go to our husband, we must attack before the wall's finished, and we must help each other is to emphasize her power and domination over her sister. She directs her even in her relation with her husband. The first-person plural pronoun 'we' which precedes the modals emphasizes solidarity, which Bodice tries to communicate to Fontanelle in order to make her sure that she seeks her interest; this in turn functions to push Fontanelle to carry out what her sister demands quite willingly. The obligation modal 'must', thus, is manipulated to channel manipulative ideology. The following frequency analysis adds more clarification for the manipulative use of the obligation modal 'must', both in the affirmative and negative forms.

TABLE V. A FREQUENCY ANALYSIS OF 'MUST'

| MUST.....TF (100)                |      |                                      |      |
|----------------------------------|------|--------------------------------------|------|
| Context                          | Word | Context                              | Line |
| (To WARRINGTON.)<br>You          | must | deal with this fever.                | 64   |
| the FIRING SQUAD).<br>They       | must | work on the wall                     | 123  |
| to help me, but you              | must | let me deal with the                 | 129  |
| But the work's slow. I           | must | do something to make                 | 136  |
| kind or merciful. I              | must | build the fortress.<br>Bodice        | 165  |
| and now you                      | must | understand ! BODICE.                 | 168  |
| Wall , wall, wall this wall      | must | be pulled down<br>Fontanelle.        | 227  |
| This is not possible! I          | must | be obeyed!<br>WARRINGTON.            | 238  |
| Fontanelle are left alone.<br>We | must | go to our husbands<br>tonight        | 322  |
| terrified of him. Bodice.<br>We  | must | attack before the wall's<br>finished | 324  |
| the Council of war. We           | must | help each other.<br>Goodbye          | 327  |
| know what she's doing! I         | must | tell her - write to                  | 2798 |
| My eyes, my eyes! I              | must | stop her before I die                | 2805 |

TABLE VI. A FREQUENCY ANALYSIS OF "MUSTN'T"

| MUSTN'T.....TF (9)    |         |                         |      |
|-----------------------|---------|-------------------------|------|
| Context               | Word    | Context                 | Line |
| this act. Bodice. You | Mustn't | talk like that in front | 151  |

Tables V and VI demonstrate that 13 occurrences of the affirmative 'must' and 1 occurrence of the negative 'mustn't' are used as carriers of manipulative ideology in the discourse of the novel. Despite its very low frequency, the negative obligation modal is highly indicative in communicating manipulative ideologies. The indication here lies in the fact that the high frequency of one word is not an indicator that this word is thematically indicative. However, low frequency words are also very indicative in many cases.

2) *Truth modality*: Many characters in Bond's play use the truth modal 'will' to reassert their trustworthiness, and to prove the validity of their speech. In Lear, the modal 'will' is used to communicate both persuasive and manipulative ideology. Here are some extract:

Fontanelle. I know you will get on with my husband. He's very understanding; he knows how to deal with old people.

Bodice. You will soon learn to respect them like your sons. (L., pp. 5-6).

Both Fontanelle and Bodice try to convince their father to bless their secret marriage from his hereditary enemies; the duke of North and the duke of Cornwall. Both of them know for sure that their father refuses their marriage, so they uses

the truth modal 'will' to influence their father's opinion towards their husbands. Fontanelle's utterance I know you'll get on with my husband signifies to state her determination to marry Cornwall. Bodice's utterance you'll soon learn to respect them like your sons is another trial to make her father bless her marriage from North. Bodice's use of the pronoun 'you' shows her power in delivering her message to her father. The two daughters try to remove the feeling of fear that occupies Lear's mind against the two husbands so as to accept their marriage without any objection. Tables VII and VIII present a frequency distribution analysis of the manipulative and persuasive 'will'.

TABLE VII. A FREQUENCY ANALYSIS OF MANIPULATIVE 'WILL'

| WILL.....TF (77)                |      |                           |      |
|---------------------------------|------|---------------------------|------|
| Context                         | Word | Context                   | Line |
| keep our enemies out.<br>People | will | live behind this wall     | 107  |
| live in peace. My wall          | will | make you free             | 109  |
| you can be -because you         | will | have my wall.             | 158  |
| LEAR. My enemies                | will | not destroy my work       | 264  |
| When I'm dead my people         | will | live in freedom and peace | 271  |

TABLE VIII. A FREQUENCY ANALYSIS OF PERSUASIVE 'WILL'

| Will.....TF (77)       |      |                            |      |
|------------------------|------|----------------------------|------|
| Context                | Word | Context                    | Line |
| Fontanelle. I know you | will | Get on with my husband     | 206  |
| Straighter! Bodice you | will | Soon learn to respect them | 213  |
| look after you. You    | will | live in decent quietness   | 3235 |
| LEAR. The wall         | will | destroy you                | 3505 |

Tables VII and VIII show that the total frequency of the truth modal 'will' is 77; only 9 occurrences are indicative in expressing particular ideologies, 5 of which are employed to convey manipulative ideology (Table VII), whereas 4 occurrences are used to communicate persuasive ideology (Table VIII). These two tables further emphasize the complementary relationship between the two variables of CATA used here: the FDA and the KWIC variables. To clarify this point, one can obviously notice that despite its ability to offer us the total frequency of a specific word, FDA still inadequate to help us better understand the indicative occurrence of a given word. Only through the variable of KWIC one can identify what is indicative and what is non-indicative among occurrences. This complementary nature of the two variables further strengthens the whole interpretative process of the analyzed text.

## V. DISCUSSION

The analysis demonstrates that pronouns and modals, sometimes, and within particular contexts, go beyond their grammatical and semantic functions to communicate and maintain particular pragmatic purposes and ideological meanings, including persuasion and manipulation. The four function words under investigation convey a specific type of ideology in the discourse of Bond's Lear. The following table adds more clarification.

TABLE IX. PRAGMATIC IDEOLOGIES OF FUNCTION WORDS IN BOND'S LEAR

| Type of function words | Linguistic manifestation | Type of ideology           |
|------------------------|--------------------------|----------------------------|
| Pronouns               | Pronoun 'I'              | persuasive<br>manipulative |
|                        | Pronoun 'we'             | manipulative               |
| Modality               | Obligation modality      | persuasive<br>manipulative |
|                        | Truth modality           | persuasive<br>manipulative |

As indicated from Table IX, communicating persuasive and manipulative ideologies is realized in Bond's Lear through pronouns, which manifest themselves in the pronouns 'I' and 'we'; each pronoun is employed to achieve a specific ideological and pragmatic purpose. The table also demonstrates that truth and obligation modalities are very indicative in communicating both persuasive and manipulative ideologies in the discourse of the play.

Other findings are also demonstrated in this study as follows:

First, the application of CATA software proves useful in extracting ideologies from language and helps better understand the power of function words, which, in turn, accentuates the analytical integration between discourse studies and computer, particularly in the linguistic analysis of large data texts. It is analytically evidenced that the two variables of FDA and KWIC are complementary in nature in the sense that the latter is a context-oriented that target the identification of indicative words generated by the total frequency analysis of the former. Significantly, both FDA and KWIC contribute to the linguistic analysis of literary texts, particularly to decipher the hidden ideologies beyond the semantic propositions of the mere linguistic expressions.

Second, the analysis demonstrated that function words, manifested here by pronouns and modality, have ideological significance in discourse. This goes in conformity with Fowler's [54] argument that there is a reciprocal relationship between language and ideology in the sense that each single linguistic unit can communicate specific ideology of its user. Ideology is usually there in language and the employment of particular linguistic expressions rather than others is ideological in nature, that is, it is produced in this particular way and in such a specific linguistic expression to communicate particular ideological meanings of the speaker/writer. Consequently, every single word can mirror the ideology of its user. In the context of this study, it is not only content words that communicate and maintain ideologies in discourse. However, function words contribute significantly in communicating and maintaining ideologies. In particular discourse contexts, function words cease to maintain their common semantic meaning to convey further ideological purposes.

Third, manipulative ideology is presented by personal pronouns ('I' and 'we'), and modality (obligation 'must' and truth 'will'). The pronouns 'I' and 'we' are used to show the

speaker's power and domination over his participants which facilitate his persuasive task. Obligation and truth modals are also used to express necessity and certitude. All these strategies are more representative in the discourse of oppression. This correlates with previous studies, such as [11], [20], and [46], which emphasize the ideological weight of pronouns and modality.

Fourth, persuasive ideology is presented in Lear through agency (pronoun I) and modality (must, will). Using agency through the pronoun 'I' reflects the desire of the speaker to express competency. Obligation and truth modals are used to express necessity and certitude, while negation is employed to expose the daughters' violence, disobedience and cruelty. These linguistic strategies serve to convey a persuasive ideology that can be said to be based on facts and past experiences. This also goes in the same direction with Sornig's [55] argument that persuasion can be realized by means of the different linguistic levels, including the grammatical one which constitutes the employment of grammatical aspects, such as the use of modality, deixis, negation, and passive structures.

## VI. CONCLUSION

This study offered a computer-aided text analysis to decode the ideological significance of function words represented by the pronouns (I, we) and the modal verbs (must, will) in the discourse of Edward Bond's Lear. The study used three analytical strands: van Dijk's ideological discourse analysis, Fairclough's model of the grammatical aspects in the analysis of discourse, and a computer-aided text analysis, which is analytically enabled by the three variables offered by CATA: frequency distribution analysis (FDA), key word in context (KWIC), and content analysis (CA). The three approaches are analytically incorporated to explore the extent to which pronouns and modality contribute to the communication of specific ideologies that vary from persuasion to manipulation in the selected text. The analysis of the selected play has evidenced the employment of pronouns and modality for ideological purposes. In terms of the use of pronouns, the analysis identified their linguistic weight as carriers of ideological agency. Sometimes, these pronouns are used inclusively, as is the case for the inclusive 'we', and in other discursive cases they are used exclusively, as is the case for the exclusive 'we'. Also, the first person singular pronoun 'I' is analytically used to communicate competency. With regard to the use of modality, the analysis showed that truth modals are used to communicate the ideological agency of certitude on the part of the user, whereas obligation modals are employed to exercise agency via expressing a high level of commitment, both on the part of speakers/writers and the hearers/readers alike. The analysis further demonstrated that pronouns and modality in the current study are employed to achieve two types of ideology: manipulative and persuasive; the former always targets the benefits of speakers/writer, while the latter often serves the benefits of all discourse participants.

This paper recommends further applications of other variables of CATA, such as LWIC (Linguistic Inquiry and Word Count) and DICTION (software package that contains

31 predefined Dictionaries) to the textual and thematic analysis of other types of function words, such as prepositions, conjunctions, and demonstratives. This might reveal different and/or similar findings than what is approached in this paper.

#### ACKNOWLEDGMENT

The researcher would like to thank Prince Sattam bin Abdulaziz University in Saudi Arabia alongside its Scientific Research Deanship for all technical support it has provided to complete this study.

#### REFERENCES

- [1] A. F. Khafaga, "A computational approach to explore the extremist ideologies of Daesh discourse". *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 193-199, 2020.
- [2] T. A. van Dijk, "Ideological discourse analysis," in E. Ventola and A. Solin (Eds.), *Interdisciplinary approaches to discourse analysis*. New Courant, 1995, pp. 135-161.
- [3] J. House, and D. Kadar, *Cross-cultural pragmatics*. Cambridge: Cambridge University Press, 2021.
- [4] T. A. van Dijk, "Discourse, knowledge and ideology: Reformulating old questions and proposing some new solutions," in P. Martin, J. N. Aertselaer, and T. A. van Dijk (Eds.), *Communicating ideologies: Multidisciplinary perspectives on language, discourse, and social practice*. New York & Oxford: Peter Lang, 2004, pp. 5-38.
- [5] R. Fowler, "On critical linguistics," in C. Coulthard, and M. Coulthard (Eds.), *Texts and practices: Readings in critical discourse analysis*. London & New York: Routledge, 1996, pp. 15-31.
- [6] A. Khafaga, "The perception of blackboard collaborate-based instruction by EFL majors/teachers amid COVID-19: A case study of Saudi universities". *Journal of Language and Linguistic Studies*, vol. 17, no. 2, pp. 1160-1173; 2021.
- [7] M. Eltahir, S. Al-Qatawneh, and S. Alsalmi, "E-Textbooks and their application levels, from the perspective of faculty members at Ajman University, U.A.E." *International Journal of Emerging Technologies in Learning*, vol. 14, no. 13, pp. 88-104, 2019.
- [8] G. Stockwell, *Computer-assisted language learning: Diversity in research and practice*. Cambridge: Cambridge University Press, 2018.
- [9] K. Beatty, *Teaching and researching computer-assisted language learning*. Harlow: Longman Pearson, 2010.
- [10] A. Khafaga, *Strategies of political persuasion in literary genres: A computational approach to critical discourse analysis*. Germany: LAMBERT Publication, 2017.
- [11] N. Fairclough, *Language and power*. London and New York: Longman, 1989.
- [12] A. F. Khafaga, "Exploring ideologies of function words in George Orwell's *Animal Farm*." *Pertanika Journal of Social Sciences and Humanities*, vol. 29, no. 3, pp. 2089 -211, 2021.
- [13] J. Charteris-Black, *Politicians and rhetoric. The persuasive power of metaphor*. Palgrave Macmillan, 2005.
- [14] A. Khafaga, and I. Shaalan, "Pronouns and modality as ideology carriers in George Orwell's *Animal Farm*: A computer-aided critical discourse analysis," *TESOL International Journal*, vol. 16, no. 4.2, pp. 78-102, 2021.
- [15] A. T. Frederiksen, and R. Mayberry, "Pronoun production and comprehension in American Sign Language: the interaction of space, grammar, and semantics". *Language, Cognition and Neuroscience*, vol. 36, no. 8, pp. 1-23, 2021.
- [16] J. Culler, *Literary theory*. Oxford University Press, 1997.
- [17] J. Culpeper, M. Short, and P. Verdonk, *Exploring the language of drama: From text to context*. London & New York: Routledge, 1998.
- [18] M. L. Heyden, J. Oehmichen, S. Nichting, and H. W. Volberda, "Board background heterogeneity and exploration-exploitation: The role of the institutionally adopted board model," *Global Strategy Journal*, vol. 5, no. 2, pp. 154-176, 2015.
- [19] D. Wiechmann, and S. Fuhs, "Concordancing software," *Corpus Linguistics and Linguistic Theory*, vol. 2, no. 2, pp. 107-127, 2006.
- [20] K. Romeo, "A web-based listening methodology for studying relative clause acquisition," *Computer Assisted Language Learning*, vol. 21, no.1, pp. 51-66, 2008.
- [21] A. Khafaga, and I. Shaalan, "Mobile learning perception in the context of COVID-19: An empirical study of Saudi EFL majors," *Asian EFL Journal*, vol. 28, no. 1.3, pp.336-356, 2021.
- [22] D. Krieger, "Corpus linguistics: What it is and how it can be applied to teaching," *The Internet TESL Journal*, vol. IX, no. 3, pp. 123-141, 2003.
- [23] J. Flowerdew, "Concordancing as a tool in course design," *System*, vol. 21, no. 2, pp. 231-244, 1993.
- [24] R. P. Weber, *Basic content analysis*. NewburyPark, CA: Sage, 1990.
- [25] I. Pollach, "Taming textual data: The contribution of corpus linguistics to computer-aided text analysis," *Organizational Research Methods*, vol. 15, no. 2, 263-287, 2012.
- [26] F. Yavuz, "The use of concordancing programs in ELT," *Procedia-Social and Behavioral Sciences*, vol. 116, pp. 2312-2315, 2014.
- [27] A. Barger, and K. Byrd, "Motivation and computer-based instructional design," *Journal of Cross-Disciplinary Perspectives in Education*, vol. 4, no. 1, pp. 1-9, 2011.
- [28] G. Fauconnier and M. Turner, "Conceptual blending, form and meaning," *Recherches en Communication [Communication Research]*, vol. 19, pp. 57-86, 2003.
- [29] T. A. van Dijk, J. Aertselaer, and M. Putz, Eds., *Introduction: Language, discourse and ideology, in Communicating ideologies: Multidisciplinary perspectives on language, discourse, and social practice*. New York & Oxford: Peter Lang, 2004, pp. xiii-xxx.
- [30] B. N. Perucha, "Ideology, cognition and discourse revisited: exploring counter-ideologies," in P. Martin, N. JoAnne, and T. A. van Dijk (Eds.), *Communicating ideologies: Multidisciplinary perspectives on language, discourse, and social practice*. New York & Oxford Peter Lang, 2004, pp. 175- 202.
- [31] A. Khafaga, "Linguistic and literary origins of critical discourse analysis." *Applied Linguistics Research Journal*, vol 5, no. 5, pp. 15-23, 2021.
- [32] T. A. van Dijk, *Discourse and knowledge: A sociocognitive approach*. Cambridge University Press, 2014.
- [33] T. A. van Dijk, "Political discourse and racism: Describing others in western parliaments," in S. H. Riggins (Ed.), *The language and politics of exclusion: Others in discourse*. Thousand Oaks, CA: Sage, 1997, , pp. 31-64.
- [34] A. F. Khafaga, "Linguistic representation of power in Edward Bond's *Lear*: A lexico-pragmatic approach to critical discourse analysis." *International Journal of English Linguistics*, vol, 9, no. 6, pp. 404-420, 2019.
- [35] R. Wodak, *Language, power and ideology: Studies in political discourse*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1989.
- [36] H. G. Widdowson, *Discourse analysis*. Oxford, Oxford University Press, 2007.
- [37] M. Thegel, and J. Lindgren, "Subjective and intersubjective modality: a quantitative approach to Spanish modal verbs." *Studia Neophilologica*, vol. 92, no. 1, pp. 124-148, 2020.
- [38] P. Dekker, "Pronouns in a pragmatic semantics. *Journal of Pragmatics*, vol. 34, no. 7, pp. 815-827, 2002.
- [39] H. Bergqvist, "Swedish modal particles as markers of engagement: Evidence from distribution and frequency." *Folia Linguistica*, vol. 54, no. 2, pp. 469-496, 2020.
- [40] D. Pinto, "Indoctrinating the youth of post-war Spain: A discourse analysis of a Fascist Civics textbook." *Discourse & Society*, vol. 15, no. 5, pp. 649-667, 2004.
- [41] R. Fowler, *Language in the news: Discourse and ideology in the press*. London: Routledge, 1991.
- [42] J. Reddington, F. Murtagh, and C. Douglas, "Computational properties of fiction writing and collaborative work." *International Symposium on Intelligent Data Analysis*, pp. 1-13, 2013.

- [43] A. Khafaga, and I. Shaalan, "Using concordance to decode the ideological weight of lexis in learning narrative literature: A computational approach," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 246-252, 2020.
- [44] A. Omar, A. F. Khafaga, and I. Shaalan, "The impact of translation software on improving the performance of translation majors," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 4, pp. 287-292, 2020.
- [45] A. Khafaga, and A. Alghawli, "The impact of CALL software on the performance of EFL students on the Saudi university context," *International Journal of Advanced Computer Science and Application*, vol. 12, no. 7, pp. 304-312, 2021.
- [46] A. Khafaga, and B. Aldossari, "The language of persuasion in courtroom discourse: A computer-aided text analysis," *International Journal of Advanced Computer Science and Application*, vol. 11, no. 7, pp. 332-340, 2021.
- [47] E. Bond, *Lear. In plays two*. London: Eyre Methuen, 1978.
- [48] A. Thabet, "Applied computational linguistics: An approach to analysis and evaluation of EFL materials." *Damietta Faculty of Education Journal*, Part 1. No. 13, pp. 7-39, 1990.
- [49] J. Sinclair, *Corpus, concordance collocation*. Oxford: Oxford University Press, 1991.
- [50] G. Kennedy, *An introduction to corpus linguistics*. London & New York: Longman, 1998.
- [51] S. Hockey, *A guide to computer applications in the humanities*. London: The Johns Hopkins University Press, 1980.
- [52] I. Inigo-Mora, "On the use of the personal pronoun 'we' in communities." *Journal of Language and Politics*, vol. 3, no. 1, pp. 27-52, 2004.
- [53] W. Abraham, *Modality in syntax, semantics and pragmatics*. Cambridge: Cambridge University Press, 2020.
- [54] R. Fowler, "Critical linguistics. In K. Malmkjar (Ed.), *The linguistic encyclopedia*. London & New York: Routledge, 1991.
- [55] K. Sornig, "Some remarks on linguistic strategies of persuasion," in R. Wodak (Ed.), *Language, power and ideology: Studies in political discourse*. John Benjamins Publishing Company, 1989, pp. 95-113.

# Fusion of BIFFOA and Adaptive Two-Phase Mutation for Helmetless Motorcyclist Detection

Sutikno<sup>1</sup>

Department of Computer Science  
Diponegoro University  
Semarang, Indonesia

Agus Harjoko<sup>2\*</sup>, Afiahayati<sup>3</sup>

Department of Computer Science and Electronics  
Universitas Gadjah Mada  
Yogyakarta, Indonesia

**Abstract**—Road traffic injuries and deaths cause considerable economic losses to individuals, families, and nations as a whole. One of the strategies needed to curtail these fatalities is the surveillance of helmetless motorcyclists, which is carried out by developing an automatic detection system based on computer vision. Generally, this system consists of three subsystems, namely, moving object segmentation, motorcycle classification, and helmetless head detection. HOPG-LDB (Histogram of Oriented Phase and Gradient - Local Difference Binary) descriptor for this system produced good accuracy; however, it still has a drawback related to a large number of features. Based on these observations, this paper proposed an Adaptive Two-phase Mutation Binary Improved Fruit Fly Optimization Algorithm (ATMBIFFOA) to reduce the features. The ATMBIFFOA is a new feature selection algorithm that improved BIFFOA (Binary Improved Fruit Fly Optimization Algorithm) with an adaptive two-phase mutation algorithm. The BIFFOA produced good accuracy; however, weak in reducing feature dimension. The adaptive two-phase mutation algorithm was used to cover this weakness. The experiment results show that the proposed method can reduce the number of features and computation time effectively from BIFFOA. The proposed method produced motorcycle classification accuracy of 96.06% for the JSC1 dataset and 96.85% for the JSC2 dataset. As for helmetless head detection, the proposed method produced an average precision of 66.29% for the JSC1 dataset and 63.95% for the JSC2 dataset.

**Keywords**—Motorcycle classification; helmetless head detection; BIFFOA; two-phase mutation algorithm

## I. INTRODUCTION

Road traffic injuries and deaths cause considerable economic losses to individuals, families, and nations as a whole. Based on the current trends, these problems are predicted to continually occur for a long period. Furthermore, World Health Organization (WHO) published that traffic accidents were the 7th leading cause of death in the world, with 1.35 million mortality cases being recorded yearly [1]. In Indonesia, the number of deaths caused by two and three-wheel motorcyclists was approximately 74% among other traffic accidents [1]. The main cause of this type of accident is the head injury sustained due to the unyieldingness of the use of helmets. WHO reported that the use of helmets reduces the risk of 69% of head injuries [1]. Most countries mandated the use of helmets; however, many motorcyclists still violate the regulation and escape the consequences, because of the difficulty of direct surveillance on the highway, which is not

monitored for a full day. Meanwhile, research in automatic detection based on computer vision has been growing rapidly, to curtail these problems.

In general, the study of detection of motorcyclists not wearing helmets was divided into two subsystems, namely motorcycle detection and helmetless head detection [2]. The feature extraction process gives an impact on the performance of both subsystems. Previous studies have used hand-crafted features and a convolutional neural network (CNN). The Histogram of Oriented Gradient (HOG) descriptor is a hand-crafted feature that results in relatively high accuracy. The author in [3] used HOG to classify vehicles in various environments and views. The author in [4] used HOG in both subsystems which resulted in good accuracy, but it still incorrectly detects distant objects. The author in [5] reported that HOG produces higher accuracy than Wavelet Transform (WT), Local Binary Pattern (LBP), and their combination in helmetless head detection. The author in [6] reported that HOG produces higher accuracy than Scale-Invariant Feature Transform (SIFT) and LBP in both subsystems.

Currently, the CNN method is popular for classification and detection in various domains. The author in [7] used CNN for motorcycle detection to overcome the problem of changing lighting and poor video quality. CNN was also used for helmetless head detection with various models, for example, AlexNet [8], VGG16 [9], VGG19, Inception V3, and MobileNets [10]. The author in [11] combined HOG and LBP for vehicle classification, and compared hand-crafted features (combination of HOG, LBP, and Haralick) and Custom CNN for helmetless head detection. The result showed that the method produces higher accuracy than HOG and LBP for motorcycle classification. For helmetless head detection, Custom CNN is superior in terms of accuracy and hand-crafted features are superior in terms of prediction time with relatively good accuracy. In addition, [12] compared several hand-crafted features (HOG, LBP, and Gabor) and CNN for vehicle classification. The result showed that the HOG produces better accuracy than other descriptors and CNN.

However, several authors stated that the HOG lacks to deal effectively with images of varying lighting [13], and different local patterns [14]. This ineffectiveness can be solved by combining HOG, Histogram of Oriented Phase (HOP), and Local Difference Binary (LDB) descriptors called Histogram of Oriented Phase and Gradient - Local Difference Binary (HOPG-LDB) [15]. The result of the experiment showed that

\*Corresponding Author

the HOPG-LDB descriptor increases the accuracy of the HOG, however, it still has a drawback related to a large number of features. The author in [16] stated that one of the preprocessing techniques that reduce these numbers is feature selection. The author in [17] stated that feature selection techniques are divided into 2, namely filters and wrappers. The author in [18] reported that the wrapper method tends to provide better performance than the filter. This technique can significantly improve the selection of relevant features [19].

The author in [20] reported that Binary Improved Fruit Fly Optimization Algorithm (BIFFOA) feature selection produces a good performance. Based on the experiment of this method compared with other algorithms, namely binary Gray Wolf Optimization (bGWO), Binary Gravitational Search Algorithm (BGSA), Binary Bat Algorithm (BBA), Binary Salp Swarm Algorithm (BSSA), Binary Grasshopper Optimization Algorithm (BGOA), Genetic Algorithm (GA), Particle Swarm Optimization (PSO), Correlation-based Feature Selection (CFS), Fast Correlation-Based Filter (FCBF), F-Score, Information Gain (IG), and spectrum. The results showed that the BIFFOA has the best accuracy, however, weak in reducing feature dimensions.

A solution for reducing feature dimensions was proposed in [21], which integrated the Gray Wolf Optimizer algorithm (GWO) and two-phase mutation (TMGWO). The first phase mutation is used to reduce feature dimensions and the second attempt to increase accuracy. The experiments of this method were compared with other algorithms, namely BBA, Binary Crow Search Algorithm (BCSA), binary Gray Wolf Optimization Algorithm (bGWOA), binary Whale Optimization Algorithm (bWOA), Discrete Particle Swarm Optimization (DPSO), Flower Algorithm (FA), Multi-Verse Optimization (MVO), PSO, and Non-Linear Particle Swarm Optimization (NLPSO). The results showed that the TMGWO produces the best accuracy and second-best feature reduction compared to other methods.

Detection of motorcyclists not wearing helmets in real-time is required the high accuracy and speed. The addition of a feature selection process can improve this performance. BIFFOA feature selection produced a good performance; however, weak in reducing feature dimension [20]. This weakness can be solved by adding a two-phase mutation algorithm. This study aims to add a feature selection process to detect motorcyclists not wearing helmets. The addition of the proposed feature selection is to reduce the number of features so the computation time of motorcycle classification and helmetless head detection can be reduced. The main contributions of this paper are:

- Adaptive Two-phase Mutation Binary Improved Fruit Fly Optimization Algorithm (ATMBIFFOA) is proposed. This algorithm is a fusion of BIFFOA and an adaptive two-phase mutation algorithm.
- An algorithm of adaptive two-phase mutation that is modified from a two-phase mutation is proposed.
- The ATMBIFFOA feature selection is added after the feature extraction process to reduce features in the motorcycle classification and helmetless head detection.

This paper is organized as follows. Related work is presented in Section II that is divided into two parts: motorcycle detection and helmetless head detection. Section III explains the dataset used, the proposed algorithm, and the evaluation methods. In Section IV, we present the experimental result and discussion. Finally, the main conclusion is introduced and future work is suggested in Section V.

## II. RELATED WORK

In general, this study is divided into two subsystems: motorcycle detection and helmetless head detection.

### A. Motorcycle Detection

Motorcycle detection has concerned three processes: vehicle segmentation, feature extraction, and classification. The author in [22] used three shape features: length, width, and their ratios to categorize vehicles into five groups. The result received from the usage of the decision tree (DT) classifier confirmed high accuracy, however, the features had been now no longer able to differentiate bicycles, motorcycles, and tricycles. The author in [23] used the features of length, width, area, diameter, and the ratio of distance to decide the object's center of mass and its main axis length. The classification method used a multilayer perceptron (MLP) to categorize the vehicles into three categories: heavy and mild duties, and motorcycles.

The author in [24] used the area feature to categorize motorcycles and others. Meanwhile, the author in [25] proposed a way that specializes in calculating the number of motorcycles on the street in real-time. The features used are area, height, and width to categorize motorcycles and non-motorcycles.

The author in [6] compared a few descriptors: HOG, SIFT, and LBP in classifying motorcycles and non-motorcycles. The effects confirmed that the HOG descriptor has exceptional accuracy. The author in [26] compared HOG, Speeded Up Robust Features (SURF), SIFT, and LBP in motorcycles detection. It has a look at extensively utilized information of images taken from in front, besides, and at the back of motorcycles. The result confirmed that the HOG descriptor has better accuracy.

An observe through as in [27] proposed a system that categorized vehicles into four categories: cars, vans, buses, and motorcycles. The system used Intensity Pyramid-based HOG (IPHOG) descriptor and support vector machine (SVM) classifier. The outcomes confirmed that the situations of climate and light converting have decreased accuracy than the normal condition.

The author in [28] used the LBP descriptor and SVM classifier to locate motorcycles. This descriptor became as compared with SURF, HOG, and Haar Wavelet. The outcomes confirmed that the proposed method has higher accuracy than the others. The author in [5] proposed a WT descriptor that became as compared with LBP, HOG, and SURF. The outcomes confirmed that the WT descriptor has higher accuracy than the others.

The author in [29] proposed a combination of shape and color features comprising of area, the ratio of width and height,

and color deviation standard. These features served as entering for the k-nearest neighbors (KNN) classifier to decide the motorcycles and non-motorcycles. The proposed approach becomes capable of calculating the wide variety of passengers on a motorcycle. The outcomes confirmed mistakes in type due to the fact the data had been taken from afar, overlapping vehicles, and the passenger sitting too near the rider. The author in [11] concatenated HOG and LBP with sequential minimum optimization (SMO) for training the SVM classifier. The outcomes display that the combination of those descriptors produced higher accuracy than the HOG and LBP descriptors. The author in [15] concatenated HOG, HOP, and LDB descriptors with MLP classifier. The results show that the proposed descriptor has higher accuracy than HOG, HOP, LDB, HOG-HOP, HOG-LDB, and HOP-LDB descriptors. Moreover, the proposed method has higher accuracy than in [5], [6], and [11].

CNN has additionally been used for motorcycle detection. The author in [30] used CNN which specializes in jam situations. The CNN is also used in [7] to address numerous lights and bad video quality. The take a look at outcomes displays that the accuracy generated is better than hand-crafted features, however, the computation time is much longer.

### B. Helmetless Head Detection

Helmetless head detection has involved three stages: ROI (region of interest) determination, feature extraction, and classification. The ROI determination pursuits to check the region round a rider's head. The heads of the rider and passenger are above the motorcycle image; therefore, the studies focused at the top a part of the object. Once the head region is known, the following steps are feature extraction and classification.

The author in [24] used the circular hough transform (CHT) descriptor for helmetless head detection. The result confirmed that it error still occurs for the detection of two or extra passengers. The author in [5] proposed the HOG descriptor, and the dataset was taken in a static environment. The assessment was performed by comparing HOG, WT, LBP, WT+LBP, WT+HOG, LBP+HOG, and WT+HOG+LBP descriptors. The result confirmed that the HOG descriptor has the best accuracy. The author in [6] compared HOG, SIFT, and LBP descriptors. The outcomes confirmed that HOG has the best accuracy. However, the data were taken on a quiet road.

Some researchers have combined shape, texture, and color features to enhance accuracy. The author in [29] used features of arc circularity, average intensity, and hue. Data were taken from three recording conditions, which include near, far, and medium. It turned into located that the greatest mistakes had been from the data recording acquired from afar.

The author in [26] used the features of arc circularity, average intensity and hue, and Center Symmetric-Local Binary Pattern (CS-LBP). These features served as entering the KNN classifier for the classification of heads with and without helmets. The technique focused on troubles with data recording

taken from special angles withinside the front, besides, and back. However, the head images were cropped manually. The author in [31] extensively utilized arc circularity, average intensity and color, and HOG.

The author in [28] used geometric, shape, and texture characteristics. The study used a combination of CHT, LBP, and HOG descriptors. CHT is used to decide the geometric form of an image. This technique has a weak point that the incapability of detecting images of low resolution. The author in [15] used the HOPG-LDB descriptor that concatenated HOG, HOP, and LDB descriptors. The results show that the HOPG-LDB descriptor has higher accuracy than HOG, HOP, LDB, HOG-HOP, HOG-LDB, and HOP-LDB descriptors. Moreover, the proposed method has higher accuracy than in [5] and [6].

The author in [32] used CNN with the YOLOv2 model to detect riders without helmets. The author in [7] used the AlexNet model on each light and heavy traffic. The author in [8] extensively utilized the AlexNet model and inaccurate detections have been made for riders placing on hats. The author in [33] used the iter\_45, Inception-V3 network, and full ImageNet network models. The author in [34] proposed Faster Regions with Convolution Neural Network (R-CNN) for decreasing the computing time. However, inaccurate detections have been nevertheless made for bicycle riders. The author in [11] compared hand-crafted features (a combination of HOG, LBP, and Haralick) and Custom CNN. Custom CNN is advanced in terms of accuracy and hand-crafted features are advanced in terms of prediction time with pretty proper accuracy.

## III. MATERIALS AND METHODS

### A. Dataset

This study used two datasets, namely JSC1 and JSC2 taken from the rear and front of an object, respectively. Both datasets contain motorcycle and non-motorcycle images used for the input of motorcycle classification. The input of helmetless head detection used the motorcycle images. Fig. 1 shows some samples of both datasets.

The datasets were generated from the segmentation process of the video. This process consists of several stages, namely histogram equalization of video frames that have been converted to grayscale, Gaussian Mixture Model (GMM) to determine foreground, and morphological operations (opening and dilation) to remove noise. The author in [35] stated that GMM is robust to lighting changes. The first video for the JSC1 dataset was recorded on Cipinang Baru Timur Street at East Jakarta, Indonesia with a frame speed of 19.49 fps. The second video for the JSC2 dataset was recorded on Budi Raya Street at West Jakarta with a frame speed of 20 fps. Both videos have a resolution of 1280 x 720 pixels and a duration of 3 hours. Training data was generated from the first 2 hours and testing data was generated from the rest. This data division technique was also used in [6]. The number of the training and testing data are shown in Table I and Table II, respectively.



Fig. 1. Samples of Datasets.

TABLE I. THE NUMBER OF TRAINING DATA

| Dataset | Motorcycle Classification Subsystem |                | Helmetless Head Detection Subsystem |                     |
|---------|-------------------------------------|----------------|-------------------------------------|---------------------|
|         | Motorcycle                          | Non-motorcycle | Head with helmet                    | Head without helmet |
| JSC1    | 1602                                | 1602           | 2052                                | 1694                |
| JSC2    | 4066                                | 4066           | 1984                                | 1984                |

TABLE II. THE NUMBER OF TESTING DATA

| Dataset | Motorcycle Classification Subsystem |                | Helmetless Head Detection Subsystem |                             |
|---------|-------------------------------------|----------------|-------------------------------------|-----------------------------|
|         | Motorcycle                          | Non-motorcycle | Motorcyclist with helmet            | Motorcyclist without helmet |
| JSC1    | 531                                 | 587            | 416                                 | 115                         |
| JSC2    | 1390                                | 1852           | 1091                                | 299                         |

### B. Developed System

In general, the system for detecting helmetless motorcyclists is divided into 3 subsystems, namely moving object segmentation, motorcycle classification, and helmetless head detection. This study focuses on developing motorcycle classification and helmetless head detection, as shown in Fig. 2. The stage of the head detection in the helmetless head detection subsystem is begun the determination process of the

ROI of the head from the motorcycle image. The ROI limits are determined based on the minimum and maximum positions of the upper 1/3 of the blob image generated from the segmentation process. The resulting image was converted to a grayscale image and was enhanced by its contrast using CLAHE (contrast-limited adaptive histogram equalization). Fig. 3(a) is an example of a motorcycle image. Fig. 3(b) is the result of this process. The next step was to create two binary images with opposite intensities using thresholding and inverse thresholding, but some blobs still need to be filtered and fixed. Fig. 3(c) shows the result of this process. The filtering was carried out by morphological operations (opening and filling holes), removing blobs on the side and top edges, and removing too big blobs. Moreover, overly tall blobs were fixed by removing the bottom. Fig. 3(d) shows the result of this process. Edge detection of Laplace of Gaussian (LoG) is used for the next step with the results as in Fig. 3(e). After that, CHT is applied to both images, and then the results are combined on an ROI head image. An example of this result is as in Fig. 3(f). The classification in the head detection is used to classify the objects bounding box on the circular into the head and non-head. Classification in helmetless head detection is used to classify head objects into heads wearing a helmet and not wearing a helmet. Fig. 3(g) is an example of the classification of head detection. The author in [5] reported that the MLP classifier produces a good performance so this paper used it. The feature extraction process used the HOPG-LDB descriptor [15].

### C. Binary Improved Fruit Fly Optimization Algorithm (BIFFOA)

The author in [20] explained that BIFFOA is developed from the Improved Fruit Fly Optimization (IFFO) algorithm for the feature selection, by converting it from continuous to binary version. The author in [36] explained that the IFFO algorithm is improved from the Fruit Fly Optimization (FFO) algorithm that is used to determine global optimization. The weakness of the FFO algorithm is that the search radius on all iterations is the same. In the IFFO Algorithm, the search radius ( $r$ ) for each iteration is calculated using (1).

$$r = r_{max} \cdot \exp\left(\log\left(\frac{r_{min}}{r_{max}}\right) \cdot \frac{Iter}{Iter_{max}}\right) \quad (1)$$

where  $r_{max}$  and  $r_{min}$  are the radii of maximum and minimum, respectively.  $Iter$  represents the iteration, and  $Iter_{max}$  represents the maximum number of iterations. The author in [37] explained that  $r_{min} = (UB-LB)/2$  and  $r_{max} = 10^{-5}$ , where  $UB$  (upper bound) and  $LB$  (lower bound) value 1 and 0, respectively.

The initialization parameters in the BIFFOA algorithm are  $PS$ ,  $r_{max}$ ,  $r_{min}$ , and  $Iter_{max}$ . In addition, the initial swarm location is initialized by selecting the best solution, which is determined by the agent with the smallest fitness value. The fitness function is designed as shown in (2).

$$Fitness = \alpha \gamma_R(D) + \beta \frac{|R|}{|C|} \quad (2)$$

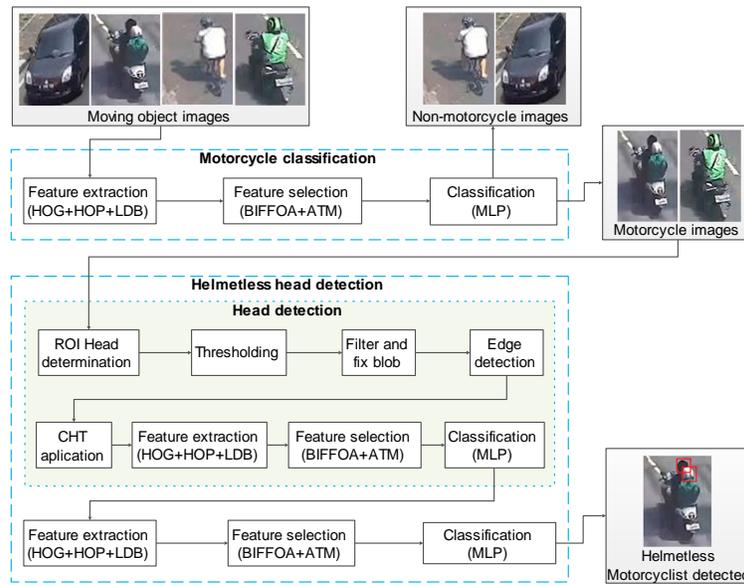


Fig. 2. A Developed System of Motorcycle Classification and Helmetless Head Detection.

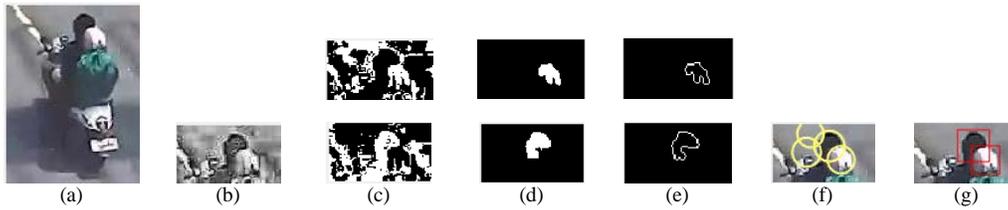


Fig. 3. Image Sample of Results from each Step: (a) Motorcycle Input (b) ROI Head Determination (c) Thresholding (d) Filter and Fix Blob (e) Edge Detection (f) CHT Application (g) Classification.

where  $\gamma_R(D)$  represents the classification error rate of a given classifier.  $|R|$  and  $|C|$  denote the length of the selected feature subset and the total number of features, respectively.  $\alpha$  and  $\beta$  represent the weight of classification accuracy and selected feature subset, respectively. The values of  $\alpha$  and  $\beta$  for this study are 0.99 and 0.01, respectively. The agents used are a swarm of fruit fly positions. The initial positions of this fruit fly are binary numbers randomly generated. The position of fruit flies is updated using (3).

$$x_{i,j} = \begin{cases} 1 - \delta_j & \text{if } S(\Delta x_{i,j}) \geq \text{rand}() \\ \delta_j & \text{otherwise, } j = 1, 2, \dots, n \end{cases} \quad (3)$$

where  $n$  is the dimension length and  $\text{rand}()$  is the generation of random numbers between  $[0, 1]$ .  $\delta_j$  is the  $j^{\text{th}}$  dimension of the optimal solution.  $S(\Delta x_{i,j})$  is the sigmoidal transfer function (S-shaped), as in (4).

$$S(\Delta x_{i,j}) = \frac{1}{1 + e^{-\Delta x_{i,j}}} \quad (4)$$

where  $\Delta x_{i,j}$  is calculated using (5).

$$\Delta x_{i,j} = \begin{cases} \delta_j \pm r \cdot \text{rand}() & \text{if } j = d \\ \delta_j & \text{otherwise, } j = 1, 2, \dots, n \end{cases} \quad (5)$$

where  $r$  is the search radius for every iteration that is calculated using equation (1).  $d$  is a dimension index that is chosen randomly. The pseudocode of the BIFFOA is shown in Algorithm 1 [20].

---

#### Algorithm 1. The standard BIFFOA

---

1. **Input:**  $PS, r_{max}, r_{min}, Iter_{max}$   
//Initialize the BIFFOA parameter:
  2. Set  $PS, r_{max}, r_{min}, Iter_{max}$
  3. Calculate the fitness of all agents using (2)
  4. Set the best solution as swarm location
  5.  $Iter=0$
  6.  $X^*=\Delta$
  7. **Repeat**
  8. Calculate the search radius  $r$  using (1)
  9. Calculate  $\Delta x_{i,j}$  using (5)  
//Ospres is foraging phase
  10. **For**  $i=1, 2, \dots, PS$
  11. Calculate the  $S(\Delta x_{i,j})$  using (4)
  12. Using (3) to generate food source,  $X_i=(x_{i,1}, x_{i,2}, \dots, x_{i,n})$
  13. **End for**  
// Vision foraging phase
  14. Calculate the fitness of all agents using (2)
  15. Update swarm location when there is a better solution in the population
  16. **Until**  $Iter=Iter_{max}$
  17. **Output:** Solution  $X^*$
- 

#### D. Two-Phase Mutation Algorithm

A two-phase mutation algorithm was proposed in [21] to improve the GWO algorithm. Algorithm 2 shows the

pseudocode of the two-phase mutation [21]. The input of this algorithm is the best grey wolf ( $X_\alpha$ ) in each iteration. The  $X_\alpha$  is mutated in two phases, the first is used to reduce features and the second is utilized in improving accuracy. The mutation is executed when the  $r$  is less than the Mutation Probability ( $M_p$ ). The value of  $r$  is between 0 and 1, which is generated randomly and the  $M_p$  value is 0.5.

---

**Algorithm 2.** The standard two-phase mutation

---

1. **Input:** the best grey wolf  $X_\alpha$  from each iteration
  2.  $Fitness =$  calculate the fitness of  $X_\alpha$   
//start the first phase
  3. Define vector *one\_positions* to store the locations of the selected features in  $X_\alpha$
  4. Define  $X_{mutated1} = X_\alpha$
  5. **For**  $i=1$  to length of *one\_positions* //for each selected feature in  $X_\alpha$
  6.     Generate a random number  $r$
  7.     **If** ( $r < M_p$ )
  8.          $X_{mutated1}[one\_positions[i]] = 0$  while keeping the other features
  9.          $Fitness\_mutated =$  the fitness of  $X_{mutated1}$
  10.        **If** ( $Fitness\_mutated < Fitness$ )
  11.            $Fitness = Fitness\_mutated$
  12.            $X_\alpha = X_{mutated1}$
  13.        **End if**
  14.     **End if**
  15. **End for**  
//start the second phase
  16. Define vector *zero\_positions* to store the locations of the unselected features in  $X_\alpha$
  17. Define  $X_{mutated2} = X_\alpha$
  18. **For**  $j=1$  to length of *zero\_positions* //for each unselected feature in  $X_\alpha$
  19.     Generate a random number  $r$
  20.     **If** ( $r < M_p$ )
  21.          $X_{mutated2}[zero\_positions[j]] = 1$  while keeping the other features
  22.          $Fitness\_mutated =$  the fitness of  $X_{mutated2}$
  23.         **If** ( $Fitness\_mutated < Fitness$ )
  24.            $Fitness = Fitness\_mutated$
  25.            $X_\alpha = X_{mutated2}$
  26.         **End if**
  27.     **End if**
  28. **End for**
  29. **Output:** the improved  $X_\alpha$
- 

**E. Proposed Algorithm: Fusion of BIFFOA with Adaptive Two-Phase Mutation Algorithm**

ATMBIFFOA is a new feature selection algorithm that is proposed in this paper. It improved the BIFFOA by adding an adaptive two-phase mutation algorithm that aims to reduce feature dimensions. The pseudocode of the ATMBIFFOA is found in Algorithm 3, while that of the adaptive two-phase mutation algorithm is found in Algorithm 4.

The input of the adaptive two-phase mutation algorithm is the best solution for each iteration ( $X^*$ ) that is defined as shown in (6).

$$X^* = (x_1, x_2, \dots, x_n) \quad (6)$$

where  $x_j$  is the  $j^{\text{th}}$  dimension of  $X^*$ , and  $n$  is the dimension length of  $X^*$ . When each  $x_j$  values= 0, then the corresponding feature is unselected. And when each  $x_j$  values= 1, then the corresponding feature is selected. The  $M_p$  is defined as the vector as shown in (7).

$$M_p = (mp_1, mp_2, \dots, mp_n) \quad (7)$$

where  $mp_j$  is the mutation probability of the  $j^{\text{th}}$  dimension. The  $mp_j$  values are constant at the beginning of iteration ( $mp_j$  in the study is 0.5). However, when the iteration is greater than or equal to the weight iteration ( $I_w$ ), then  $M_p$  is equal to the dimension weights of the best agents. The  $I_w$  is calculated using (8).

$$I_w = t_w \times Iter_{max} \quad (8)$$

where  $t_w$  is a weight threshold that values a range of [0, 1].

The dimension weights of the best agents are represented in the vector ( $W_i$ ), as in (9).

$$W_i = (w_{i,1}, w_{i,2}, \dots, w_{i,n}) \quad (9)$$

where  $w_{i,j}$  is the weight of the best agent in the  $i^{\text{th}}$  iteration and  $j^{\text{th}}$  dimension that is calculated using (10).

$$w_{i,j} = \frac{w_{(i-1),j} \times (i-1) + x_{i,j}}{i} \quad (10)$$

where  $x_{i,j}$  is the value of the best solution in the  $i^{\text{th}}$  iteration and  $j^{\text{th}}$  dimension.

---

**Algorithm 3.** The proposed ATMBIFFOA

---

1. **Input:**  $PS, r_{max}, r_{min}, Iter_{max}$   
//Initialize the ATMBIFFOA parameter:
  2. Set  $PS, r_{max}, r_{min}, Iter_{max}$
  3. Set  $W_0, M_p, I_w$
  4. Calculate the fitness of all agents using (2)
  5. Set the best solution as swarm location
  6.  $Iter = 0$
  7.  $X^* = \Delta$
  8. **Repeat**
  9. Calculate the search radius  $r$  using (1)
  10. Calculate  $\Delta x_{ij}$  using (5)  
//Ospres is foraging phase
  11.     **For**  $i=1, 2, \dots, PS$
  12.         Calculate  $S(\Delta x_{ij})$  using (4)
  13.         Using (3) to generate food source,  $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$
  14.     **End for**  
// Vision foraging phase
  15. Calculate the fitness of all agents using (2)
  16. Update the swarm location when there is a better solution in the population
  17. Update  $W_i$  using (9)
  18.     **If**  $Iter \geq I_w$
  19.          $M_p = W$
  20.     **End if**  
//Mutation process
  21. Process of the adaptive two-phase mutation
  22. **Until**  $Iter = Iter_{max}$
  23. **Output:** Solution  $X^*$
-

The mutation process in the two-phase mutation algorithm is executed in all dimensions of one\_position and zero\_position vector. Therefore, this algorithm takes a long time when used in a large number of features. The adaptive two-phase mutation algorithm limited the number of mutated dimensions with the 1<sup>st</sup> Mutation Candidate Probability ( $P_{mc1}$ ) and the 2<sup>nd</sup> Mutation Candidate Probability ( $P_{mc2}$ ), therefore, its computation time can be reduced. The mutation position of both vectors is selected through random permutation.

**Algorithm 4.** The proposed adaptive two-phase mutation

1. **Input:** the best solution  $X^*$  from each iteration  
//start the first phase
2. Define vector *one\_positions* to store the locations of the selected features in  $X^*$
3. Define  $X_{mutated1} = X^*$
4. Define the number of mutation candidate  $n_{mc} = P_{mc1} \times$  length of *one\_position*
5. Define vector *one\_mutation\_candidate* to store the location of the mutated candidate by selecting  $n_{mc}$  random permutations in *one\_position*.
6. **For**  $i=1$  to length of *one\_mutation\_candidate*
7.     Generate a random number  $r$
8.     **If** ( $r < M_p[\text{one\_mutation\_candidate}[i]]$ )
9.          $X_{mutated1}[\text{one\_mutation\_candidate}[i]] = 0$   
while keeping the other features
10.         *Fitness\_mutated* = the fitness of  $X_{mutated1}$
11.         **If** (*Fitness\_mutated* < *Fitness*)
12.             *Fitness* = *Fitness\_mutated*
13.              $X^* = X_{mutated1}$
14.     **End if**
15.     **End if**
16. **End for**  
//start the second phase
17. Define vector *zero\_positions* to store the locations of the unselected features in  $X^*$
18. Define  $X_{mutated2} = X^*$
19. Define the number of mutation candidate  $n_{mc} = P_{mc2} \times$  length of *zero\_position*
20. Define vector *zero\_mutation\_candidate* to store the location of the mutated candidate, by selecting the  $n_{mc}$  random permutations in the *zero\_position*
21. **For**  $j=1$  to the length of *zero\_mutation\_candidate*
22.     Generate a random number  $r$
23.     **If** ( $r < M_p[\text{zero\_mutation\_candidate}[j]]$ )
24.          $X_{mutated2}[\text{zero\_mutation\_candidate}[j]] = 1$   
while keeping other features
25.         *Fitness\_mutated* = the fitness of  $X_{mutated2}$
26.         **If** (*Fitness\_mutated* < *Fitness*)
27.             *Fitness* = *Fitness\_mutated*
28.              $X^* = X_{mutated2}$
29.     **End if**
30.     **End if**
31. **End for**
32. **Output:** the improved  $X^*$

**F. Evaluation Methods**

The parameters for measuring performance were feature number (NF), time of average classification (Time), accuracy (Acc), precision (Pre), and recall (Rec). Especially for helmetless head detection, we used average precision (AP) to measure accuracy. Equations (11), (12), and (13) are used in calculating accuracy, precision, and recall, respectively.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$Recall = \frac{TP}{TP+FN} \quad (13)$$

where TP, TN, FP, and FN represent true positive, true negative, false positive, and false negative, respectively. The TP is the true detection of the ground-truth bounding box. The correct detection was measured using the intersection over union (IOU) as in (14).

$$IOU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \quad (14)$$

where  $B_p$  and  $B_{gt}$  represent the predicted and ground-truth bounding box, respectively. The detection is considered correct when the IOU is greater than or equal to the threshold. In this study, the threshold value was 0.5.

AP is the area under the precision-recall curve which has a range of [0, 1] and it is calculated as in (15) [38].

$$AP = \sum_n (R_{n+1} - R_n) P_i(R_{n+1}), \quad (15)$$

where  $P_i(R_{n+1})$  is calculated using (16).

$$P_i(R_{n+1}) = \max_{\tilde{R}: \tilde{R} \geq R_{n+1}} P(\tilde{R}) \quad (16)$$

where  $P(\tilde{R})$  is the precision measured at the time of recall  $\tilde{R}$ .

The experiment was carried out by selecting the best result on each process, namely a combination of cell and block sizes on the HOPG-LDB descriptor, variation of  $t_w$  value on the ATMBIFFOA, and a combination of hidden layer number, neuron number, and training algorithm on the MLP. The block size variations were 2x2 and 3x3 cells. The cell size variations in the 2x2 blocks were 4x4, 6x6, 8x8, and 12x12 pixels and the 3x3 block sizes were 4x4 and 8x8 pixels. The variations of  $t_w$  values were 0.25, 0.5, and 0.75. The variations of the number of hidden layers used are 1, 2, and 3. The number of neurons in the hidden layers ( $n_H$ ) was determined by using (17) [39].

$$n_H = \sqrt{n_i + n_o} + l \quad (17)$$

where  $n_i$  is the number of neurons in the input layer,  $n_o$  is the number of neurons in the output layer, and  $l$  is an integer constant of 1 to 10. The  $l$  variation of this study was 1, 5, and 10. Finally, we used 8 variations of the training algorithm, namely the gradient descent with adaptive learning rate backpropagation (traingda), scaled conjugate gradient backpropagation (trainscg), conjugate gradient backpropagation with Powell-Beale restarts (traincgb), conjugate gradient backpropagation with Fletcher-Reeves update (traincgf), conjugate gradient backpropagation with Polak-Ribière update (traincgp), one step secant backpropagation (trainoss), gradient descent with momentum and adaptive learning rate backpropagation (traingdx), and gradient descent backpropagation (traingd). This paper used the learning rate of 0.05, the epoch maximum number of 1000, and the limit for error of 0.001 for the training. The author in [40] reported that these parameters result in a good performance.

For the ATMBIFFOA, the parameters of PS and  $Iter_{max}$  are 24 and 100, respectively. The values of  $P_{mc1}$  and  $P_{mc2}$  are 0.25 and 0.01, respectively. The K-fold cross-validation (K=5) was used to separate the training and validation data in the feature selection process. Each experiment is run 5 times and the average results are used. All the experiments were carried out in Windows 10 Ultimate 64-bit operating system, with processor Intel Core (TM) i7-9750HQ CPU and 16 GB of RAM. All the algorithms were implemented in the MATLAB R2019a Software.

#### IV. RESULTS AND DISCUSSION

This section shows the experiment results of the proposed method for motorcycle classification and helmetless head detection.

##### A. Motorcycle Classification

The first experiment is to determine the best accuracy of the proposed method (ATMBIFFOA) with variations of  $t_w$ . Table III shows the results of this experiment for the motorcycle classification. From this table, we can be seen that the best accuracy reaches 96.06% for the JSC1 dataset and 96.85% for the JSC2 dataset.

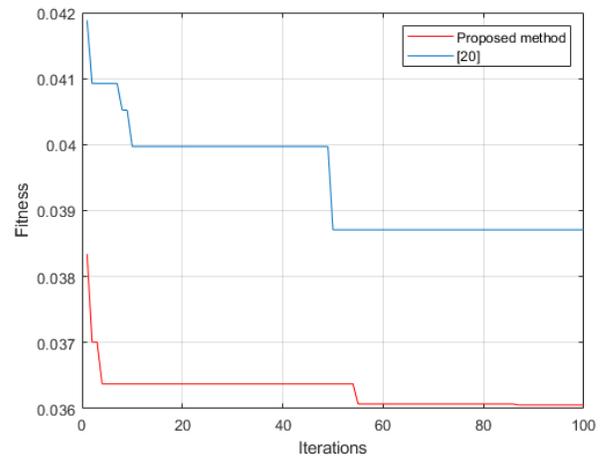
Furthermore, the proposed method is compared with the previous study, namely in Table IV. Here, [20] used BIFFOA feature selection. From this table, it can be seen that the proposed method is superior in terms of the number of features and classification time. Meanwhile, in terms of accuracy, the proposed method is superior for the JSC1 dataset, and [20] is superior for the JSC2 dataset. For this reason, we conclude that the addition of an adaptive two-phase mutation algorithm in BIFFOA can effectively reduce the number of features.

TABLE III. EXPERIMENTAL RESULTS WITH VARIATION OF  $t_w$  FOR MOTORCYCLE CLASSIFICATION

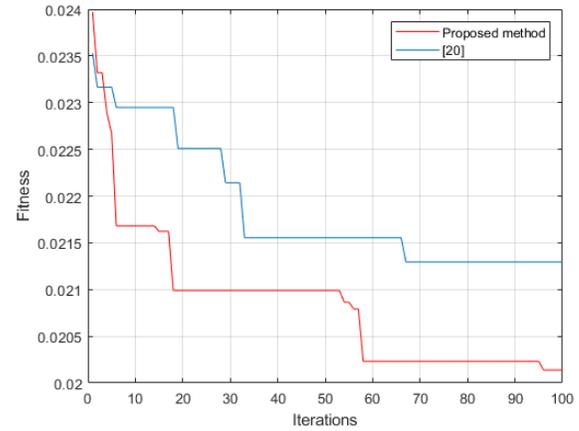
| Dataset | $t_w$ | Acc (%)      | Pre (%)      | Rec (%)      | NF           | Time ( $\times 10^{-3}$ s) |
|---------|-------|--------------|--------------|--------------|--------------|----------------------------|
| JSC1    | 0.25  | 95.90        | 95.38        | 96.05        | 899.0        | 46.4321                    |
|         | 0.50  | <b>96.06</b> | <b>95.63</b> | <b>96.12</b> | 951.8        | 42.6672                    |
|         | 0.75  | 95.39        | 95.28        | 95.03        | <b>829.6</b> | <b>38.1412</b>             |
| JSC2    | 0.25  | 96.71        | 97.28        | 94.98        | 232.8        | 34.9727                    |
|         | 0.50  | <b>96.85</b> | 97.38        | <b>95.21</b> | <b>227.8</b> | <b>34.2448</b>             |
|         | 0.75  | 96.71        | <b>97.46</b> | 94.81        | 253.0        | 41.5804                    |

TABLE IV. FEATURE SELECTION COMPARISON BETWEEN PROPOSED METHOD AND PREVIOUS STUDY

| Dataset | Method          | NF           | Time ( $\times 10^{-3}$ s) | Acc (%)      | Pre (%)      | Rec (%)      |
|---------|-----------------|--------------|----------------------------|--------------|--------------|--------------|
| JSC1    | [20]            | 1155.4       | 48.685                     | 96.05        | 95.62        | 96.08        |
|         | Proposed method | <b>951.8</b> | <b>42.667</b>              | <b>96.06</b> | <b>95.63</b> | <b>96.12</b> |
| JSC2    | [20]            | 283.6        | 42.320                     | <b>96.96</b> | <b>97.67</b> | 95.19        |
|         | Proposed method | <b>227.8</b> | <b>34.244</b>              | 96.85        | 97.38        | <b>95.21</b> |



(a) JSC1 Dataset.



(b) JSC2 Dataset.

Fig. 4. Convergence Curve of the Proposed Method and the Previous Study for Motorcycle Classification.

Fig. 4 displays the convergence curve of the proposed method and the previous study in [20]. The proposed method produces a better optimal solution than the BIFFOA.

##### B. Helmetless Head Detection

Table V shows the experimental results of helmetless head detection with variations in the value of  $t_w$ . From this table, it can be seen that the highest AP reaches 66.29% for the JSC1 dataset and 63.95% for the JSC2 dataset. Furthermore, the proposed method is compared with previous studies.

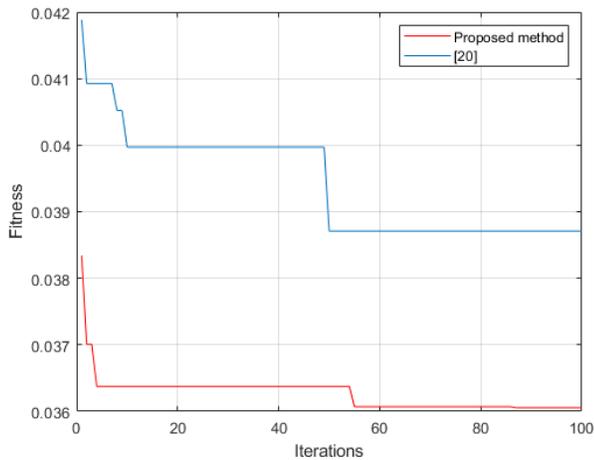
Table VI shows a comparison of the proposed feature selection method and previous study. Here, [20] used the BIFFOA feature selection. From this table, it can be seen that the proposed method is superior in terms of the number of features and classification time. In addition, the AP of the proposed method is superior for the JSC2 dataset, although it is slightly lower for the JSC1 dataset. Fig. 5 shows the comparison of the convergence curve between the proposed method and the previous study. The proposed method produces a better optimal solution than [20]. Therefore, we can conclude that the addition of an adaptive two-phase mutation algorithm to BIFFOA can reduce features effectively.

TABLE V. EXPERIMENTAL RESULTS WITH A VARIETY OF  $t_w$  FOR HELMETLESS HEAD DETECTION

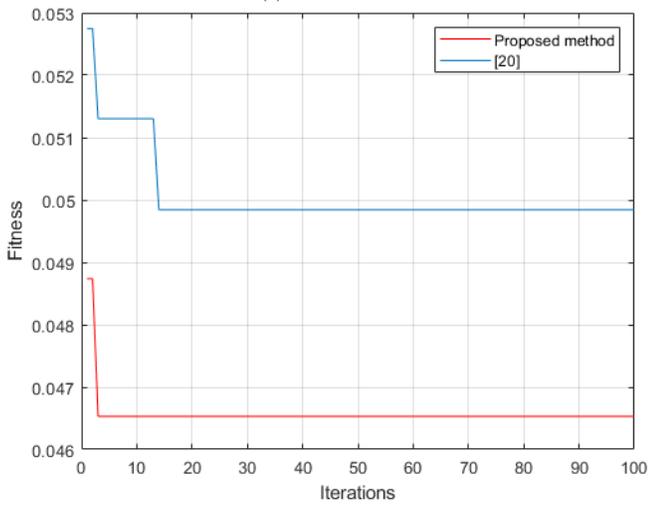
| Dataset | $t_w$ | AP (%)       | Pre (%)      | Rec (%)      | NF           | Time ( $\times 10^{-3}$ s) |
|---------|-------|--------------|--------------|--------------|--------------|----------------------------|
| JSC1    | 0.25  | <b>66.29</b> | <b>57.19</b> | <b>76.74</b> | <b>138.6</b> | 4.489                      |
|         | 0.50  | 63.59        | 55.63        | 75.56        | 151.6        | 4.499                      |
|         | 0.75  | 66.09        | 54.90        | <b>76.74</b> | 147.6        | <b>4.487</b>               |
| JSC2    | 0.25  | 60.49        | 51.65        | 79.18        | 416.2        | 6.044                      |
|         | 0.50  | <b>63.95</b> | <b>52.68</b> | <b>81.82</b> | <b>391.2</b> | <b>4.464</b>               |
|         | 0.75  | 61.92        | 49.98        | 80.95        | 419.0        | 5.929                      |

TABLE VI. FEATURE SELECTION COMPARISON BETWEEN PROPOSED METHOD AND PREVIOUS STUDY FOR HELMETLESS HEAD DETECTION

| Dataset | Method          | NF           | Time ( $\times 10^{-3}$ s) | AP (%)       | Pre (%)      | Rec (%)      |
|---------|-----------------|--------------|----------------------------|--------------|--------------|--------------|
| JSC1    | [20]            | 159.6        | 4.641                      | <b>66.68</b> | <b>57.31</b> | <b>77.19</b> |
|         | Proposed method | <b>138.6</b> | <b>4.489</b>               | 66.29        | 57.19        | 76.74        |
| JSC2    | [20]            | 445.0        | 5.255                      | 62.55        | 51.59        | 81.23        |
|         | Proposed method | <b>391.2</b> | <b>4.464</b>               | <b>63.95</b> | <b>52.68</b> | <b>81.82</b> |



(a) JSC1 dataset



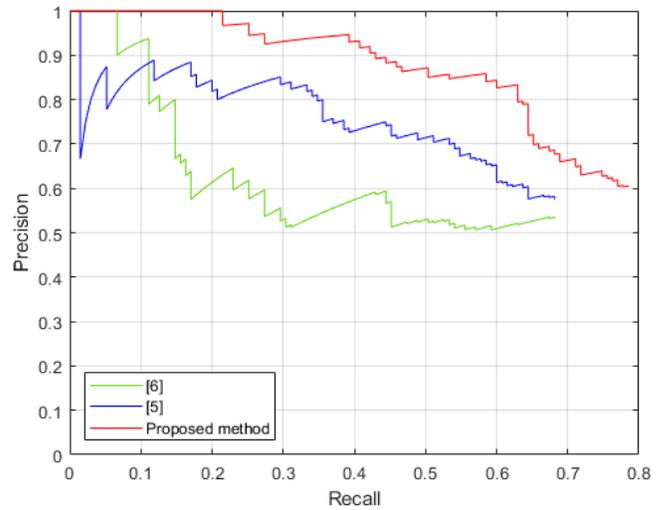
(b) JSC2 dataset

Fig. 5. Convergence Curve of the Proposed Method and the Previous Study for Helmetless Head Detection.

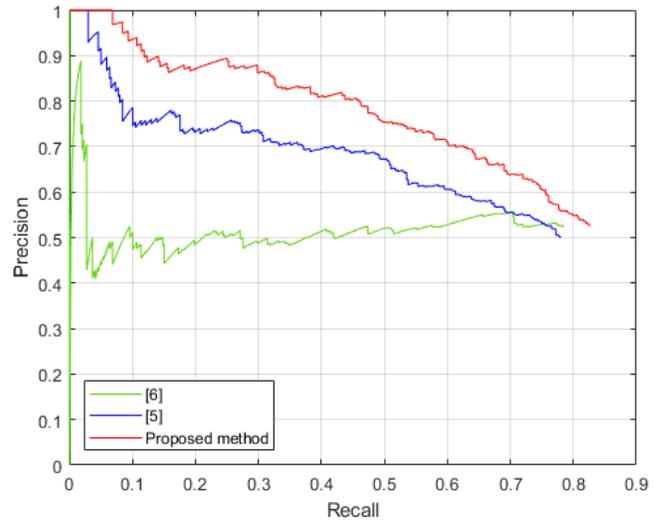
The proposed method is also compared with previous studies, as shown in Table VII. Here, [6] used a combination of HOG descriptor and SVM classifier, and [5] used a combination of HOG descriptor and MLP classifier. AP of the proposed method is superior when compared to these methods. Fig. 6 shows a comparison of the precision-recall curve of the proposed method and these methods.

TABLE VII. COMPARISON BETWEEN PROPOSED METHOD AND PREVIOUS STUDIES FOR HELMETLESS HEAD DETECTION

| Dataset | Method          | AP (%)       | Pre (%)      | Rec (%)      |
|---------|-----------------|--------------|--------------|--------------|
| JSC1    | [6]             | 43.41        | 53.18        | 68.15        |
|         | [5]             | 52.45        | <b>57.50</b> | 68.15        |
|         | Proposed method | <b>66.29</b> | 57.19        | <b>76.74</b> |
| JSC2    | [6]             | 40.47        | 52.27        | 78.59        |
|         | [5]             | 54.30        | 50.00        | 78.13        |
|         | Proposed method | <b>63.95</b> | <b>52.68</b> | <b>81.82</b> |



(a) JSC1 dataset



(b) JSC2 dataset

Fig. 6. The Curve of Precision-Recall of the Proposed Method and the Previous Study.

## V. CONCLUSION

This paper proposed a new feature selection algorithm called ATMBIFFOA for motorcycle classification and helmetless head detection. The experiment used two datasets with different recording angles, namely the rear and front of an object. The motorcycle classification accuracy of the proposed method reaches 96.06% for the JSC1 dataset and 96.85% for the JSC2 dataset. Meanwhile, the AP of helmetless head detection reaches 66.29% for the JSC1 dataset and 63.95% for the JSC2 dataset. The proposed algorithm is more effective than BIFFOA in terms of the number of features and the time of classification. For this reason, the proposed method is more suitable for the detection of motorcyclists who do not wear helmets in real-time. However, the addition of an adaptive two-phase mutation algorithm to BIFFOA can significantly increase the feature selection time. In the future, ATMBIFFOA can be used with faster classifiers such as KNN, SVM, and DT to reduce the time consumption of feature selection.

## ACKNOWLEDGMENT

The authors would like to thank the Research Directorate of Universitas Gadjah Mada in the Recognisi Tugas Akhir (RTA) 2021 schema for funding this research.

## REFERENCES

- [1] W. H. O. (WHO), "Global Status Report on Road Safety 2018," Available at: <https://apps.who.int/iris/rest/bitstreams/1164010/retrieve>, March 2021.
- [2] Sutikno, Afiahayati, and A. Harjoko, "Detection on head of motorcyclist without helmet: a review," *ICIC Express Letters*, Vol. 14, pp. 605-612, June 2020.
- [3] A. S. Azim, A. Jafri, and A. Alkhaury, "Multiclass vehicle classification across different environments," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 12, pp. 681-691, March 2021.
- [4] V. L. Padmini, G. K. Kishore, P. Durgamalleswarao, and P. T. Sree, "Real time automatic detection of motorcyclists with and without a safety helmet," *Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020)*, Trichy, India, pp. 1251-1256, September 2020.
- [5] R. R. V. E. Silva, K. R. T. Aires, and R. M. S. Veras, "Detection of helmets on motorcyclists," *Multimedia Tools and Applications*, Vol. 77, pp. 5659-5683, March 2018.
- [6] K. Dahiya, D. Singh, and C. K. Mohan, "Automatic detection of bike-riders without helmet using surveillance videos in real-time," *Proceeding of International Joint Conference on Neural Networks (IJCNN)*, Vancouver, Canada, pp. 3046-3051, July 2016.
- [7] C. Visnu, D. Singh, C. K. Mohan, and S. Babu, "Detection of motorcyclists without helmet in videos using convolutional neural network," *Proceedings of International Joint Conf. on Neural Networks (IJCNN)*, Anchorage, USA, pp. 3036-3041, May 2017.
- [8] K. C. D. Raj, A. Chairat, V. Timtong, M. N. Dailey, and M. Ekpanyapong, "Helmet violation processing using deep learning," *Proceedings of International Workshop on Advanced Image Technology (IWAIT)*, Chiang Mai, Thailand, pp. 1-4, January 2018.
- [9] Y. Kulkarni, S. Bodkhe, A. Kamte, and A. Patil, "Automatic number plate recognition for motorcyclists riding without helmet," *Proceedings of International Conference on Current Trends toward Converging Technologies*, Coimbatore, India, pp. 1-6, March 2018.
- [10] N. Boonsirisumpun, W. Puarungroj, and P. Wairotchanaphuttha, "Automatic detector for bikers with no helmet using deep learning," *Proceedings of International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai, Thailand, pp. 1-4, November 2018.
- [11] L. Shine and C.V. Jiji, "Automated detection of helmet on motorcyclists from traffic surveillance videos: a comparative analysis using hand-crafted features and CNN," *Multimedia Tools and Applications*, Vol. 79, pp. 14179-14199, February 2020.
- [12] S. Baghdadi and N. Aboutabit, "View-independent vehicle category classification system," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 12, pp. 756-771, July 2021.
- [13] H. K. Ragb and V. K. Asari, "Histogram of oriented phase and gradient (HOPG) descriptor for improved pedestrian detection," *Proceedings of the IS & T International Conf. on Electronic Imaging: Video Surveillance and Transportation Imaging Applications*, San Francisco, USA, pp. 1-6, February 2016.
- [14] H. Wang, D. Zhang, and Z. Miao, "Fusion of LDB and HOG for face recognition," *Proceedings of the 37th Chinese Control Conf. (CCC)*, Wuhan, China, pp. 9192-9196, July 2018.
- [15] Sutikno, A. Harjoko, and Afiahayati, "Improving detection performance of helmetless motorcyclists using the combination of HOG, HOP, and LDB descriptors," *International Journal of Intelligent Engineering & System*, Vol. 15, pp. 428-440, February 2022.
- [16] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, Vol. 112, pp. 103375 (1-9), September 2019.
- [17] Z. Zhu, Y. Ong, and M. Dash, "Wrapper-filter feature selection algorithm using a memetic framework," *IEEE Transactions on Systems*, Vol. 37, pp. 70-76, February 2007.
- [18] J. Zhang, Y. Xiong, and S. Min, "A new hybrid filter/wrapper algorithm for feature selection in classification," *Analytica Chimica Acta*, Vol. 1080, pp. 43-54, November 2019.
- [19] K. Bouzoubaa, Y. Taher, and B. Nsiri, "Predicting DOS-DDOS attacks: review and evaluation study of feature selection methods based on wrapper process," *International Journal of Advanced Computer Science and Applications (IJACSA)*, Vol. 12, pp. 131-145, May 2021.
- [20] Y. Hou, J. Li, H. Yu, and Z. Li, "BIFFOA : a novel binary improved fruit fly algorithm for feature selection," *IEEE Access*, Vol. 7, pp. 781177-81194, July 2019.
- [21] M. Abdel-Basset, D. El-Shahat, I. El-Henawy, V. H. C. Albuquerque, and S. Mirjalili, "A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection," *Expert Systems with Applications*, Vol. 139, pp. 1-14, January 2020.
- [22] A. Leelasantham and W. Wongsere, "Detection and classification of moving thai vehicles based on traffic engineering knowledge," *Proceedings of International Conference on ITS Telecommunication*, Phuket, Thailand, pp. 27-30, October 2008.
- [23] S. Fazli, S. Muhammadi, and M. Rahmani, "Neural network based vehicle classification for intelligent traffic control," *International Journal of Software Engineering & Applications*, Vol. 3, pp. 17-22, May 2012.
- [24] T. Marayatr and P. Kumhon, "Motorcyclist's helmet wearing detection using image processing," *Advanced Materials Research*, Vol. 931-932, pp. 588-592, May 2014.
- [25] Y. Dupuis, P. Subirats, and P. Vasseur, "Robust image segmentation for overhead real-time motorbike counting," *Proceedings of International IEEE Conf. on Intelligent Transportation Systems (ITSC)*, Qingdao, China, pp. 3070-3075, October 2014.
- [26] M. Ashvini, G. Revathi, B. Yogameena, and S. Saravanaperumaal, "View invariant motorcycle detection for helmet wear analysis in intelligent traffic surveillance," *Proceedings of International Conf. on Computer Vision and Image Processing*, Roorkee, India, pp. 175-185, December 2016.
- [27] Z. Chen, T. Ellis, and S.A. Velastion, "Vehicle detection, tracking and classification in urban traffic," *Proceedings of International IEEE Conf. on Intelligent Transportation Systems*, Anchorage, USA, pp. 951-956, September 2012.
- [28] R. Silva, K. Aires, T. Santos, K. Abdala, R. Veras, and A. Soares, "Automatic detection of motorcyclists without helmet," *Proceedings of Latin America Computing Conf. (CLEI)*, Caracas, Venezuela, pp. 1-7, October 2013.
- [29] R. Waranusast, N. Bundon, and P. Pattanaburt, "Machine vision techniques for motorcycle safety helmet detection," *Proceedings of International Conf. on Image and Vision Computing New Zealand*, Wellington, New Zealand, pp. 35-40, November 2013.

- [30] C. Huynh, T. Le, and K. Hamamoto, "Convolutional neural network for motorbike detection in dense traffic," Proceedings of the International Conf. on Communications and Electronics (ICCE), Ha-Long, Vietnam, pp. 369-374, July 2016.
- [31] A.S. Talautikar, S. Sanathanan, and C.N. Modi, "An enhanced approach for detecting helmet on motorcyclists using image processing and machine learning techniques," Advanced Computing and Communication Technologies, Vol. 702, pp 109-119, July 2019.
- [32] J. Mistry, A.K. Misraa, M. Agarwal, A. Vyas, V.M. Chudasama, and K.P. Upla, "An automatic detection of helmeted and non-helmeted motorcyclist with license plate extraction using convolutional neural network," Proceedings of Seventh International Conf. on Image Processing Theory, Tools and Applications (IPTA), Montreal, Canada, pp. 1-6, November 2017.
- [33] M. A. V. Forero, "Detection of motorcycles and use of safety helmets with an algorithm using image processing techniques and artificial intelligence models," Proceedings of Joint Conf. for Urban Mobility in the Smart City, Medellin, Colombia, pp. 1-9, April 2018.
- [34] V. Mayya and A. Nayak, "Traffic surveillance video summarization for detecting traffic rules violators using R-CNN," Proceedings of International Conf. on Computer, Communication, and Computational Sciences, Kathu, Thailand, pp. 117-126, October 2017.
- [35] Y. Xu, J. Dong, B. Zhang, and D. Xu, "Background modeling methods in video analysis: a review and comparative evaluation." CAAI Transactions on Intelligence Technology, Vol. 1, pp. 43-60, January 2016.
- [36] W. Pan, "A new fruit fly optimization algorithm: taking the financial distress model as an example," Knowledge-Based Systems, Vol. 26, pp. 69-74, February 2012.
- [37] S. Mousavi, N. Alikar, and S. Niaki, "An improved fruit fly optimization algorithm to solve the homogeneous fuzzy series-parallel redundancy allocation problem under discount strategies," Soft Computing, Vol. 20, pp. 2281-2307, June 2016.
- [38] R. Padilla, S. L. Netto, and E. A. B. D. Silva, "A survey on performance metrics for object-detection algorithms," Proceedings of the International Conference on Systems, Signals, and Image Processing (IWSSIP), Niteroi, Brazil, pp. 237-242, July 2020.
- [39] X. Chen, Z. Liu, and Z. Zhang, "The measurement of planning surface roughness by neural networks based on image," Proceedings of Sixth International Conference on Natural Computation, Yantai, China, pp. 705-708, August 2010.
- [40] H. Mustafidah, S. Hartati, R. Wardoyo, and A. Harjoko, "Selection of most appropriate backpropagation training algorithm in data pattern recognition," International Journal of Computer Trends and Technology (IJCTT), Vol. 14, pp. 92-95, September 2014.

# AI-based System for the Detection and Prevention of COVID-19

Sofien Chokri<sup>1</sup>, Wided Ben Daoud<sup>2</sup>, Wasma Hanini<sup>3</sup>, Sami Mahfoudhi<sup>4</sup>, Amel Makhoulf<sup>5</sup>

NTS'Com Research Unit, ENET'COM, University of Sfax, Sfax, Tunisia<sup>1,2,5</sup>

Laboratory of the Advanced Electronic Systems and the Durable Energy (ESSE), University of Sfax, Sfax, Tunisia<sup>3</sup>

Department of Management Information Systems and Production Management, College of Business and Economics, Qassim University, Buraydah 52571, Saudi Arabia<sup>4</sup>

**Abstract**—The COVID-19 pandemic has had catastrophic consequences all over the world since the detection of the first case in December 2019. Currently, exponential growth is expected. In order to stop the spread of this pandemic, it is necessary to respect sanitary protocols such as the mandatory wearing of masks. In this research paper, we present an affordable artificial intelligence-based solution to increase the protection against COVID-19, covering several relevant aspects to facilitate the detection and prevention of this pandemic: non-contact temperature measurement, mask detection, automatic gel-dispensing, and automatic sterilization. Our main contribution is to provide high-quality, real-time learning and analysis. To achieve this goal, we used a deep convolutional neural network (CNN) based on MobileNetV2 architecture as the learning algorithm and Advanced Encryption Standard (AES) as an encryption protocol for sending secure data to notify hospital staff. The experimental results show the effectiveness of our model by providing 99.7% accuracy in detecting masks with a runtime of 1.54 s.

**Keywords**—Face mask detection; coronavirus; COVID-19; deep learning; MobileNetV2; AES

## I. INTRODUCTION

The world faces a serious pandemic named COVID-19 as a result of the new SARS-CoV-2 virus, which began in China in late December 2019. This epidemic spread quickly beyond China on February 25, 2020 for the first time [1]. By January 20, 2022, the total number of identified cases was 338 807 207 while 5 581 841 individuals had passed away worldwide [2]. The most common symptoms of COVID-19 are fever, headache and loss of smell, muscle pains, and dry cough. In the most severe forms, the onset of acute respiratory distress syndrome leads to death, especially in people who are more fragile because of their age or comorbidities [3]. Coronavirus infections are frequent in humans and in most cases, they are transmitted directly (from one person to another) by respiratory droplets. However, they are also transmitted indirectly by surfaces [4]. To reduce the spread of this disease, many protective and safety measures have been taken by the authorities such as mandatory wearing of an indoor mask, physical distancing, self-isolation, limitation of citizen movement within a country's borders and abroad, closure of non-essential workplaces and educational institutions, and finally reduction of public transport and restriction of domestic and international travel [5]. Overall, the sanitary protocols of

preventing COVID-19 have shown positive results in reducing the spread of the virus [6].

Consequently, we propose a new system of access control in compliance with health protocols. In this article, we present an intelligent system to help organizations comply with COVID-19 security rules and reduce the spread of the pandemic. We focus on the most common internal measures, such as the distance between people, which should be at least 1.5 to 2 meters. Equally crucial is wearing a mask and washing hands with hydro alcoholic gel. Finally, people with a temperature above 38°C should stay at home and receive health care, and should not go to work or school or interact with others outside the home.

Our system is designed to help the fight against this pandemic by monitoring the wearing of masks to reduce the spread of viruses [7]. Moreover, this system detects the fever of people without making contact. The room is sterilized daily and automatically when an abnormal temperature is detected. Gel is distributed without the need to touch the dispenser since COVID-19 could be transmitted by a plastic surface. Finally, the algorithm used in this work detects mask-wearing at 99.7% efficiency.

The remainder of this paper is organized as follows: The second section explores some related works and similar approaches to COVID-19 detection. In the third section, we outline our suggested method for COVID-19 detection and prevention. Then, in the fourth section, we describe the materials and methods used in the experiment, including the results of the evaluation. In the fifth section, we offer the conclusion.

## II. RELATED WORK

It has become increasingly important to consider some methods of COVID-19 prevention and detection to limit and restrict the rapid spread of the epidemic virus. In their works, several researchers have proposed numerous approaches to detect COVID-19 and protect persons from this virus.

The authors in [8] have proposed a portable non-contact method to screen the health status of people wearing masks through analysis of respiratory characteristics. In their work, they proposed the use of a device which consists of a thermal camera FLIR ONE and an Android phone. This work is based on face detection in a video stream to capture technical breathing data. Then, the deep learning algorithm Gated

Recurrent Unit (GRU) is applied to the respiratory data to obtain the result of medical screening. While this method can help to combat the current epidemic of COVID-19, the proposed algorithm is not stable in the respiratory status measurement because of the effects of different types of masks..

In another work [9], the authors focus on the workflow based on the detection of COVID-19 from image classification using the deep learning model's convolutional neural network (CNN), they provided a pre-processing pipeline aimed at removing the sampling bias and improving the image quality. The results show that the CNN algorithm based on a Visual Geometry Group (VGG19) model provides better COVID-19 detection results against pneumonia for the ultrasound images. This work enables quick, accessible, affordable, and reliable identification of COVID-19 and helps to slow the transmission of COVID-19 infection. However, the used database requires a large number of images for better identification.

Jordi Laguarda and all, in [10] built a data collection pipeline of COVID-19 cough recordings between April and May 2020. They created the largest balanced COVID-19 audio cough to develop an intelligent speech-processing framework that leverages acoustic biomarker feature extractors to pre-screen COVID-19 from cough recordings.

In [11], the authors proposed a mask-wearing detection system. This system recognizes whether or not a person is wearing a mask based on the transfer learning technique. The ResNet-50 model is based on YOLO v2 (deep learning algorithm), which gives an accuracy of 81% of mask detection. The major drawback of this work is that the authors did not give details about the confusion matrix of the proposed algorithm.

Regarding temperature detection, there are several variant Arduino-based solutions. For instance, in [12], IoT systems for security monitoring based on temperature detection were presented. However, the used sensor does not give accurate results. In this context, in [13], the authors present a comparison between three types of sensors to measure temperature without making contact. According to the obtained results, the thermal camera Lepton gives the most accurate value with a minimum error rate.

To summarize, the previously cited works were carried out to design and deploy COVID prevention and detection systems. However, we deduce that these systems perform poorly for small and unvaried datasets because of the great variety of masks and the differing of symptoms from one person to another.

To remedy the problems mentioned above, our main objective is to offer a COVID-19 detection and prevention system based on deep learning for the detection of mask-wearing. We provide an effective and intelligent solution with a secure monitoring process based on the MobileNetV2. Moreover, because of the sensitivity of transmitted results to doctors, we need to ensure confidentiality. In fact, to provide secure end-to-end service, data are encrypted based on symmetric cryptography using AES (advanced encryption algorithm).

### III. OVERVIEW OF THE PROPOSED SCHEME

The pandemic of COVID-19 is spreading rapidly every day. This disease has become a major threat to people's lives as it often causes death. Therefore, we base our research on the prevention and detection of this disease in the field of work and the health sector.

In this section, we provide an overview of our proposal to achieve the previously presented goals. We start by defining the system model containing the elements of our solution, where we implement specific algorithms. Indeed, we present an intelligent solution for the detection of and prevention from coronavirus. Our proposed model consists of the following subsystems, illustrated in Fig. 1.

Our model is useful in buildings such as hospitals, clinics, and company headquarters. Indeed, we contribute the design of an intelligent system to detect and prevent COVID-19 based on the use of deep learning algorithms to classify people with and without mask, and an algorithm for non-contact temperature detection. Our solution is based on the extraction of significant face parameters such as eyes and nose, followed by the application of the CNN algorithm based on MobileNetV2 architecture. Then, to detect the temperature, we apply an algorithm that generates a thermal image and measures the temperature. After detecting an abnormal temperature value, we use an encryption algorithm to send the result to the hospital. These results are decrypted using an open-source application. Furthermore, to clean hands when entering the building, we use a system that allows the automatic contactless distribution of alcohol gel.

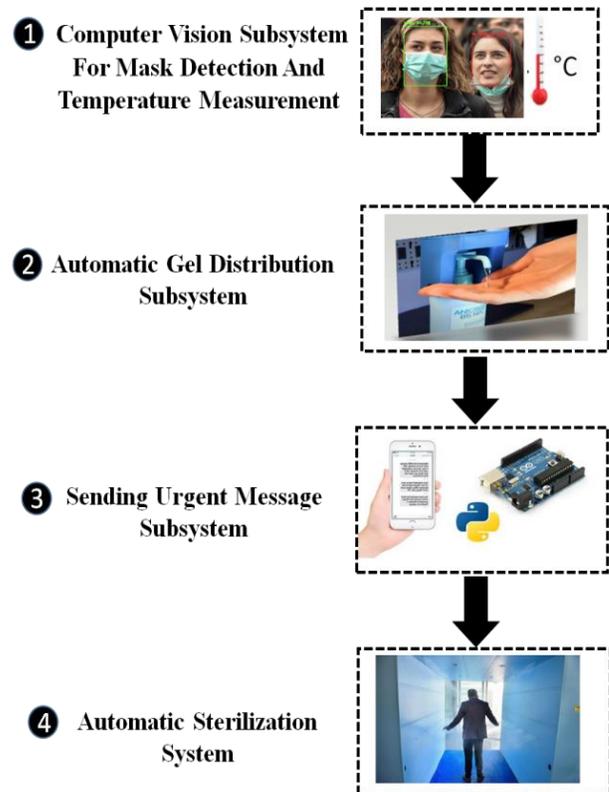


Fig. 1. COVID-19 Detection and Prevention Model.

Finally, for the automatic sterilization phase, we ensure our system is equipped with high quality cleaning products to sterilize the room every day and following the detection of abnormal temperature.

Fig. 2 presents a flowchart of the COVID-19 detection and prevention process, beginning with the presence of a person. First, he/she must practice social distancing, which at minimum is equal to 1.5 meters. After that, she/he moves on to the next step of mask detection.

If the person is not wearing a mask or not covering their nose, an audible signal is sent to warn him to wear the mask correctly and repeat the test. Otherwise, he/she must leave. If the mask is worn correctly, the second step is processed to measure the temperature without contact. For this task, the thermal sensor will be triggered using an infrared sensor to measure the temperature.

In the event that this person has a body temperature above normal, the door remains closed and an audible signal informs this person to go to the waiting room for five minutes and then repeat a second temperature measurement test. For this duration of waiting, we used a timer which is triggered by a sensor after entering the waiting room. After the second test, if the person's temperature remains high, a signal is displayed informing them to exit and wait for the ambulance service. As soon as the sensor detects the exit of a person, an audible message informs the other persons to leave. After three minutes, the sterilization is triggered automatically using a cleaning product against the virus. However, if the person is wearing the mask correctly and the temperature does not exceed the normal degree, then the door opens and it goes to the gel-dispensing stage. For this step, if the sensor detects the presence of a hand at a distance of less than 7 cm, the gel is dispensed automatically. Otherwise, if the distance is greater than 7 cm, the system remains closed.

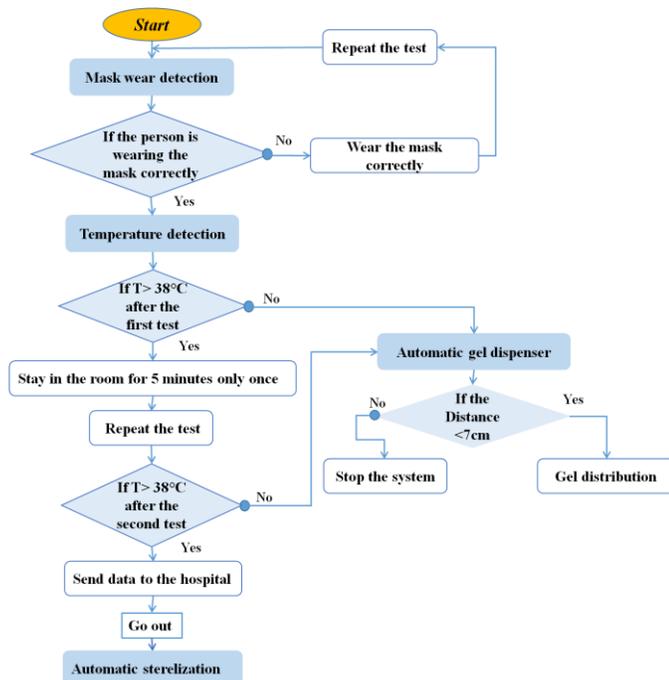


Fig. 2. The Proposed Detection and Prevention Model.

### A. Mask Detection Algorithm

For our mask detection algorithm, we rely on MobileNetV2, which is a convolution optimization for convolutional neural networks. This small, low latency, low power model is adjusted to meet the resource constraints of a variety of use cases. MobileNetV2 is a highly efficient architecture that can be applied to embedded devices with limited computing capacity. Used for feature extraction, facial recognition, and object detection, it is based on separable in-depth convolution as the base unit. This convolution has two layers: deep convolution and point convolution (1 \* 1 convolution) [14].

An inverted residual structure is used to allow the network to calculate the activations (ReLU) more efficiently, and to retain more information after activation. These connections are between the bottleneck layers.

MobileNetV2 architecture contains the fully convolutional initial layer with 32 filters, followed by 19 bottleneck residual layers [15]. Algorithm 1 presents the steps of the proposed system.

---

#### Algorithm 1: mask port detection

---

**Inputs:** database containing different images with and without masks

**Outputs:** categorized images showing the presence of a face mask

For the images in the database of two categories,

1. Convert RGB (red, green, blue) images into grayscale images
2. Resize the images to 224\*224
3. Normalize the image and convert it to a four-dimensional array

**End**

#### To build the MobileNetV2 model,

1. Add a convolution layer of 32 filters
2. Add a convolution layer 1\*1
3. Insert a flattening layer in the network classifier
4. Add a dense layer to activate ReLU
5. Add AveragePooling2D
6. Add the final dense layer with two outputs for two categories

**End**

#### Train the model

---

We use OpenCV (Open Source Computer Vision Library) to detect faces in an image [16]. This software is a very popular algorithm that is used to detect one or more faces in the image (Fig. 3).

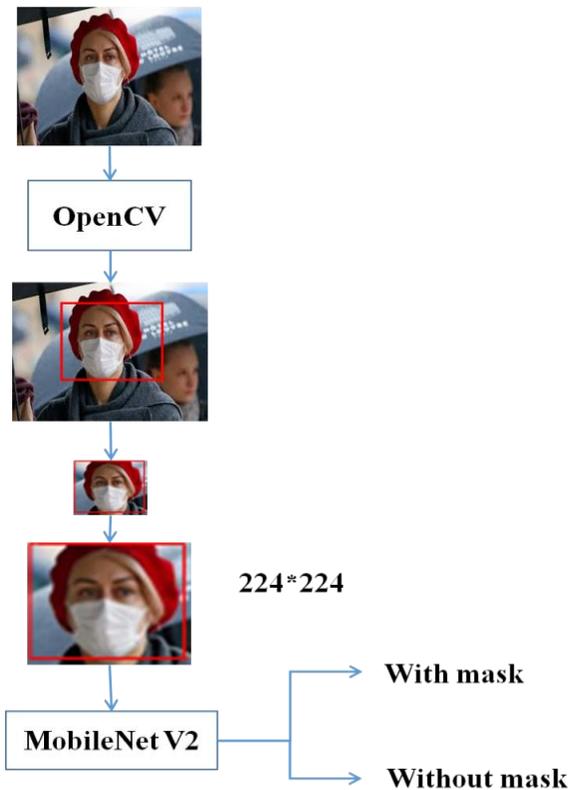


Fig. 3. Mask Detection Chain.

### B. Fever Detection Module

For the detection of fever, we propose the temperature detection process following Algorithm 2.

---

#### Algorithm 2: Fever detection algorithm

---

**Input:** Amount of infrared energy emitted by an object

**Output:** Read the temperature of the sensor

1. Measure the amount of infrared energy
2. Calculate the signal using DSP (calculation unit)
3. Convert to a temperature value using an ADC
4. Generate data via the I2C communication protocol
5. Read the temperature

If  $t > 38^{\circ}\text{C}$ , the door remains closed and the person enters the waiting room

Otherwise, if  $t < 38^{\circ}\text{C}$ , the door is opened automatically

**End**

---

The flowchart illustrated by Fig. 4 summarizes the mask and temperature detection. When the person wears the mask properly and his temperature does not exceed  $38^{\circ}\text{C}$ , the door opens automatically followed by the automatic alcohol gel-dispensing step to clean hands from viruses. Alternatively, if the sensor detects an abnormal temperature with a false mask port, then the door remains closed.

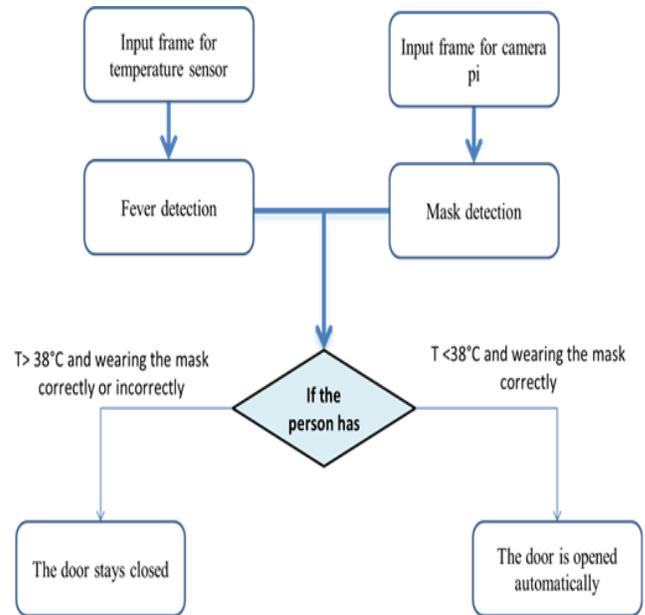


Fig. 4. Flowchart for Mask and Temperature Detection.

### C. Automatic Gel Dispenser

For the gel dispenser, we use a specific sensor that measures the distance between a body and the used component. As presented above, if a person places their hand less than 7 cm in front of the sensor, the gel will be automatically dispensed without touching it. We use this contactless system to clean the person's hands from viruses before entering the building to reduce the transmission of this epidemic.

### D. Message-Sending System

For this task, after detecting the temperature using the thermal sensor and verifying that it exceeds normal temperature, a message is sent quickly to the hospital. This message contains the location of the company headquarters and the degree of detected abnormal temperature. To secure the sending of the message against malicious attacks, our system is based on the AES-CBC-256 algorithm. AES guarantees that sensitive data is only accessible for authorized users to read.

### E. Automatic Sterilization System

To curb the spread of this epidemic, our system is based on high quality cleaning products. Therefore, the automatic sterilization system uses a water pump to dispense the cleaning products against the virus. The purpose of sterilization is to reduce the population of microorganisms, facilitate cleaning, protect personnel when handling instruments, and avoid contamination of the environment.

## IV. EXPERIMENTS AND RESULTS

We describe in this section the detailed realization of our proposed system. Fig. 5 introduces the architecture of the detection and prevention system, subdivided into four subsystems. For the mask port and temperature detection system, we chose a Raspberry Pi 4 model b card. This card is equipped with a 1.5 GHz processor and 8 Go RAM. We chose a USB-type camera to allow better resolution of 8 megapixels for mask detection against a 5 megapixel camera. For

contactless fever detection, we chose MLX90614 thermal sensor for its advantages of low cost and small size. As previously described, when the system has checked that the person is wearing the mask correctly and his temperature does not exceed the normal degree, then the door opens, and an SG 90 servo motor is used to switch to the automatic gel distribution system. To this end, we have chosen the SG 90 servo motor as an ultrasonic sensor and an Arduino Nano board equipped with a microprocessor. For the detection of the body, we propose using ATmega328, and for the message-sending phase, we chose an Arduino mega 2560 card equipped with a microprocessor: ATmega2560 and a GSM SIM 800L V2 module. The GSM module SIM800L V2 starts and searches the network automatically, and has low energy consumption. It can be directly connected to Arduino which has 5V level [17]. If the module receives a signal via serial communication from the USB port with the Raspberry Pi, it sends directly to the hospital an encrypted message containing the location of the company headquarters and the abnormal temperature of the person. This encryption is determined by the AES algorithm. For the automatic sterilization system, we used a speaker, a mini water pump, an ultrasonic sensor, and an Arduino Nano board.

The experiments are conducted in the Anaconda environment (version 5.2.0) using Python language. The experimental configuration computer is an Intel i5-3317U processor at 1.70 GHz with 6 GB of RAM. We present in the following section the results of each module.

#### A. Mask Detection Camera

1) *MobileNetV2 implementation:* To develop the model, we must first import the required functions from the Keras ML library. Keras is a deep learning API written in Python. Keras enables rapid experimentation which is able to move from

idea to result as quickly as possible. The basic data structures of Keras are layers and models [18]. All layers used in the MobileNetV2 model are implemented using Keras.

2) *Database description:* The dataset [19] consists of 3833 images in which 1915 images are people who are wearing face masks and the remaining 1918 images are people who are not wearing face masks. This large number of images is used to train and test our algorithm to improve the performance of the model. Fig. 6 contains mainly a front face pose with different mask colors.

First, each image is converted from RGB (red, green, blue) to a grayscale image that contains a single-color channel (the “grayscale” of each pixel). The images are then resized to reduce the complexity and computational power of the MobileNetV2 model.

In this algorithm, we used the cv2.cvtColor function to convert the RGB image to grayscale, and the cv2.resize function to resize our image to the dimensions (img\_size, img\_size). Our img\_size parameter was set to 224, so that each image becomes a 224\*224 square image. The prepared image is added to the ‘data’ list and the class label is added to the ‘labels’ list. To optimize the training time and reduce the complexity of the model, we convert the ‘data’ list into a more efficient NumPy array, and then we divide the array by 255, which normalizes the pixel range between 0 and 1. In the proposed model, we used TensorFlow to reshape the data (image) in data processing.

TensorFlow is an open-source interface developed by Google researchers to perform deep learning and other statistical and predictive analysis workloads. It is designed for running advanced analytics applications for users such as predictive modelers and data scientists [20, 21].

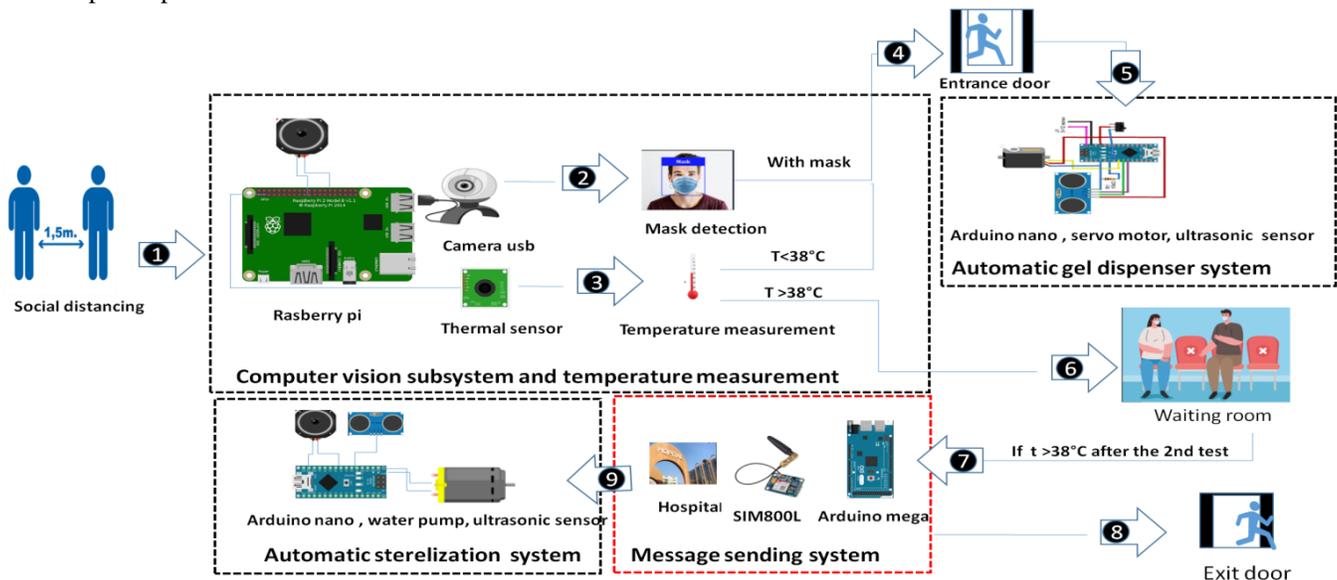


Fig. 5. Proposed Evaluation Scheme.



Fig. 6. Image Data Set with Two Classes of “with Mask” and “without Mask”.

3) *Simulation results:* The evaluation of our proposal is based on different metrics, where the confusion matrix, the false positive (FP), the true positive (VP), and the true negative (TN) rates are calculated. Fig. 7 presents the confusion matrix. This matrix is in the form of a table which is often used to describe the performance of our classification model on the set of images with and without masks. Our algorithm detects with a true positive 1569 out of 1574 images with masks, and detects with a true negative 1462 out of 1574 images without masks.

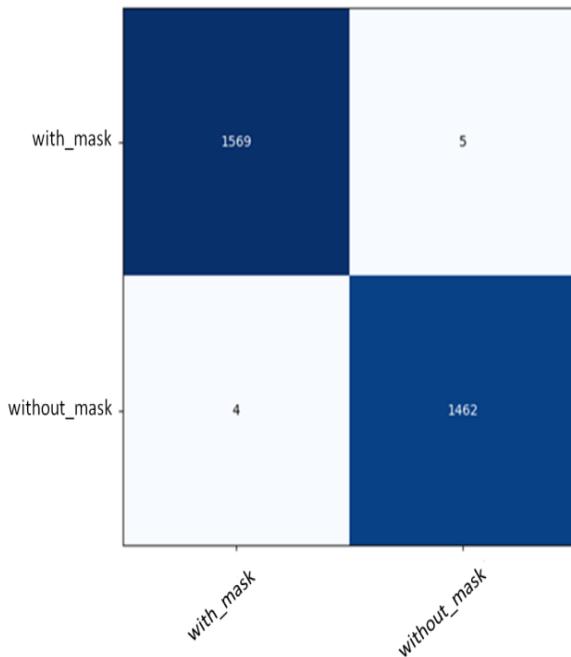


Fig. 7. Confusion Matrix.

The first intuitive indicator of success is precision. It is the quantitative relationship between correctly predicted positive observations and total predicted positive observations [22].

The formula (1) is used to calculate the precision of our algorithm, where our method achieves an accuracy of up to 99.7%.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} * 100 \quad (1)$$

Moreover, Recall or Sensitivity is the quantitative relationship of positive observations correctly predicted, or all observations within the actual class-yes [23]. It is calculated following Equation (2). Our model reaches a Recall equal to 99.68%.

$$\text{Recall} = \frac{TP}{TP + FN} * 100 \quad (2)$$

The F1 score (Equation (3)) is used to assess a two-class system. It is a method which combines the precision of the model and the recall rate. It is defined as the harmonic mean of the model accuracy and the recall rate, where the result retrieved from our model is. F1 score = 99.69%.

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Accuracy})}{(\text{Recall} + \text{Accuracy})} \quad (3)$$

Fig. 8 illustrates the contrast between loss of training and corresponding validation to the dataset. One of the main reasons for obtaining this precision resides in Average Pooling [24].

A much higher number of neurons and filters can cause a decrease in performance. The optimized values of the filters and the size of the pool allow the main part (face) of the image to be filtered in order to detect the presence of a face.

The system is able to effectively detect faces which are partially obscured either with a mask, hair, or a hand. It considers the degree of occlusion over four regions –nose, mouth, chin, and eyes – to differentiate an annotated mask from a face covered by a hand.

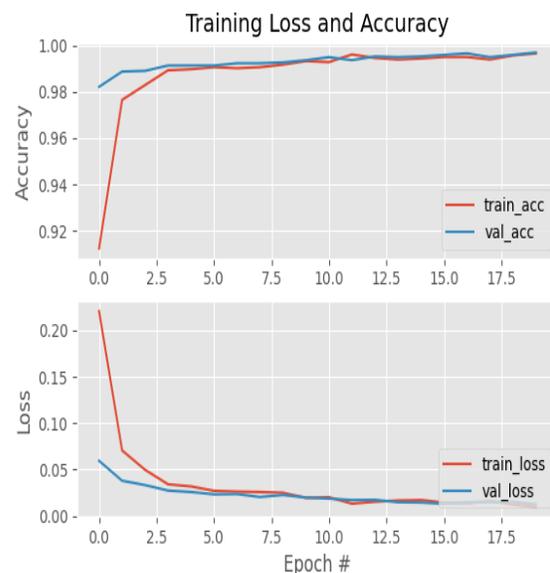


Fig. 8. Training Loss and Accuracy of the Model.

MobileNetV2 works best for all types of masks, thereby introducing the effectiveness of the proposed model in detecting masked faces. We improve our model using the Adam optimizer. Table I illustrates the results of comparing different related work methods in terms of precision. The work presented in [12] used the Medical Masks Dataset (MMD). The authors of [11] obtained an average test precision equal to 81% using YOLO v2 with ResNet-50, where in [25] the authors use the CNN model, which achieves a validation precision of 96% for the detection of facial mask. In [26], the authors use Real-World Masked Face Recognition Dataset (RMFRD) in their work. They obtained a test accuracy equal to 97% using ResNet-50.

By analyzing the performance of MobileNetV2 in the management of all types of masks, we find that our model is the most efficient and fast, where it achieves a mask detection accuracy of 99.7% thanks to its optimized architecture that contains inverted residuals and linear bottlenecks.

TABLE I. COMPARISON OF RESULTS

| References | Algorithms             | Precision in classifying images |
|------------|------------------------|---------------------------------|
| [11]       | YOLO v2 with ResNet-50 | 81%                             |
| [25]       | CNN                    | 96%                             |
| [26]       | ResNet50               | 97%                             |
| Our work   | MobileNetV2            | 99,7%                           |

On the webcam stream (as shown in Fig. 9) of using the MobileNetV2, the classification results are displayed on a label above the visual rectangle.



Fig. 9. Results obtained without Masks and with Masks.

### B. Fever Detection Module

For fever detection, we used MLX90614 non-contact sensor [27]. This infrared (IR) temperature sensor can be used to measure the temperature of particular objects ranging from -70°C to 380°C. It includes two built-in devices: one is infrared thermopile detector (detection unit) and the other is DSP signal conditioning device (computing unit). The sensor uses IR rays to measure the temperature of the object without making any physical contact, and it communicates with the microcontroller using the I2C protocol. The sensor measures both the object temperature and the ambient temperature to calibrate the object temperature value. The detection results are shown in Fig. 10. As evident in this figure, if the infrared sensor detects the presence of the hand, then the MLX90614 will automatically measure the temperature and the result obtained below the visual rectangle.



Fig. 10. Simulation Result for Non-Contact Temperature Measurement.

We conducted the test on 130 students from our school by using an IR thermometer with the MLX90614. Table II summarizes our assessment. The results show that there is an average difference between the two sensors of 2.5°C for measuring a person's temperature. Our system only has two errors of incorrect temperature measurement. To conclude, our system helps us to reduce the spread of this pandemic.

TABLE II. EVALUATION OF OUR PROTOTYPE

|                                                              |                           |
|--------------------------------------------------------------|---------------------------|
| <b>Number of students tested</b>                             | 130                       |
| <b>Test result for MLX90614</b>                              | Between 32.6°C and 34.3°C |
| <b>IR thermometer test result</b>                            | Between 35.1°C and 36.8°C |
| <b>Temperature degree difference between mlx90614 and IR</b> | 2.5°C                     |
| <b>Number of errors</b>                                      | 2                         |
| <b>Test duration for each student</b>                        | 2.6 seconds               |
| <b>CPU temperature for the Raspberry board</b>               | Between 42°C and 59°C     |
| <b>Detection distance for two sensors</b>                    | 2cm                       |

Fig. 11 illustrates the functional prototype for the detection of mask wear and fever.

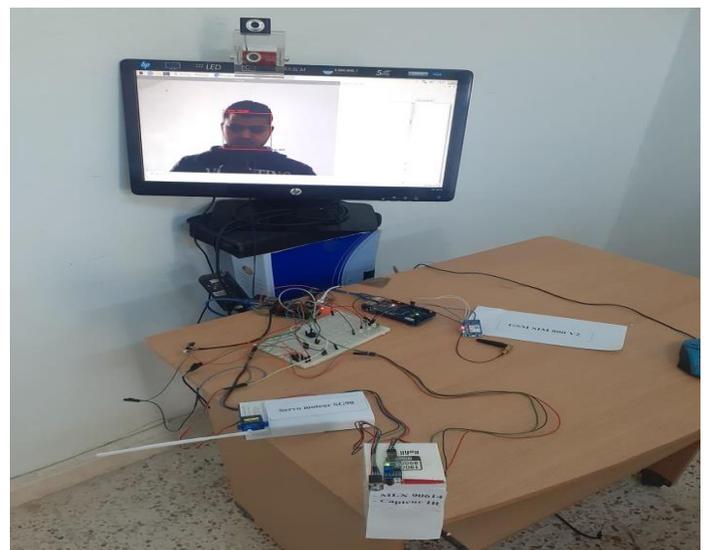


Fig. 11. Functional Prototype for the Detection of Mask Wear and Fever.

### C. Automatic Gel Dispenser

For this module, we used the ultrasonic sensor. We send a high pulse of  $10\mu\text{s}$  to the trigger pin of the sensor. Then, this sensor sends a series of eight ultrasonic pulses at 40 KHz (inaudible to the human ear). The ultrasounds propagate in the air until touching an obstacle and then return in the other direction towards the sensor. After that the sensor detects the echo and triggers the measurement. Finally, the signal on the echo pin of the sensor remains high. This allows for the duration of the round trip of the ultrasound to be measured and thus the distance to be determined [28]. The Equation (4) calculates the distance.

$$\text{Distance} = (\text{pulse duration (in } \mu\text{s)} / 2) / 29.1 \quad (4)$$

Using Proteus 8 Professional software, we obtained the results presented in Fig. 12. In fact, if the ultrasonic sensor detects the presence of a body, then the servo motor turns to distribute the alcohol gel.

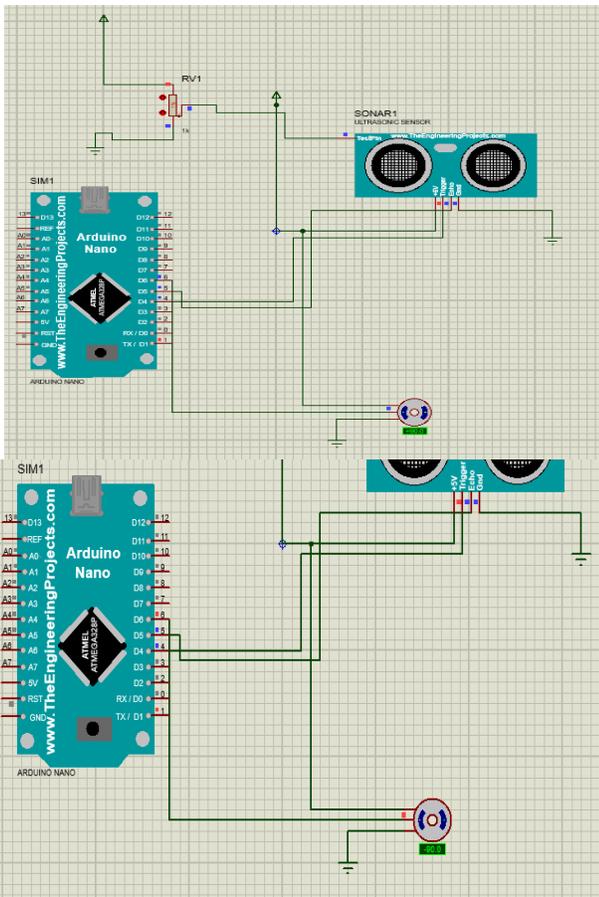


Fig. 12. Functional Schemefor the Automatic Gel Distribution.

Fig. 13 shows the experimental result for the automatic dispensing of gel. The sensor detects the presence of a human hand to disperse the gel automatically.



Fig. 13. Experimental Scheme for the Automatic Gel Distribution.

### D. Message-Sending System

After an abnormal temperature detection, the system sends a signal to the message-sending system via USB communication. This message is based on Arduino Mega 2560 board and GSM SIM 800LV2 module. If this signal is received, then the module has to send an encrypted message quickly to the hospital from the SIM card. This message contains the location of the company headquarters and the abnormal temperature level of the person.

For the decryption phase, we used an open-source application called CrypTool [29]. Its use is simple: when receiving the encrypted message, the receiver must put it in the field message encrypt with the secret code of 24 bit. Then, the message will be decrypted automatically after 1s.

### E. Automatic Sterilization System

To reduce the transmission of COVID-19 [30], we have built a smart contactless system to automatically sterilize the room. In this regard, our system is based on ultrasonic sensor, water pump, speaker, push button, and Arduino Nano. Fig. 14 show the experimental results.

In fact, after that the ultrasonic sensor detects the exit of the person, the sound message will be triggered to inform other people to exit. After three minutes, the Arduino board will send a signal to the water pump to automatically dispense the cleaning product. For the sterilization phase each day, we have provided our system with a push button to control the start of cleaning product dispensing.

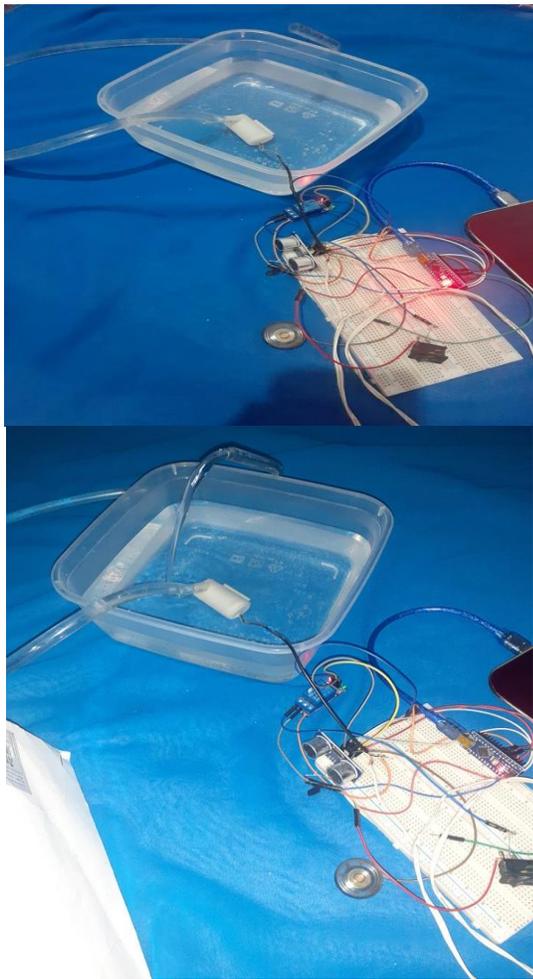


Fig. 14. Experimental Scheme of the Automatic Sterilization System.

## V. CONCLUSION

In this article, we present an intelligent temperature and mask detection system. The main objective is to improve internal security against COVID-19 by ensuring a detection and prevention method.

To achieve this goal, we applied a deep convolutional neural network based on the MobileNetV2 architecture. We used a real database composed of masked and unmasked images. Moreover, we use AES-CBC- 256 as the encryption protocol to ensure security and privacy information users. The MLX90614 is implemented as a contactless temperature detection sensor. The results show that by using OpenCV, Keras, and TensorFlow, our MobileNetV2 architecture operates with a maximum accuracy of 99.7% and a runtime of 1.53s. The maximum contactless temperature detection distance must not exceed 2cm.

For future work, we will improve our algorithm to detect false mask-wearing, to identify building personnel wearing their masks, and record their temperature every day.

## ACKNOWLEDGMENTS

This work was supported by the Deanship of Scientific Research, Qassim University.

## REFERENCES

- [1] Mohan, B. S.; Vinod, Nambiar (2020) COVID-19. An Insight into SARS-CoV2 Pandemic Originated at Wuhan City in Hubei Province of China. In : Journal of Infectious Diseases and Epidemiology, vol. 6, n° 4. DOI: 10.23937/2474-3658/1510146.
- [2] COVID-19 CORONAVIRUS PANDEMIC: <https://www.worldometers.info/coronavirus/>.
- [3] Sanyaolu, Adekunle; Okorie, Chuku; Marinkovic, Aleksandra; Patidar, Risha; Younis, Kokab; Desai, Priyank et al. (2020) Comorbidity and its Impact on Patients with COVID-19. In : SN comprehensive clinical medicine, p. 1–8. DOI: 10.1007/s42399-020-00363-4.
- [4] Karia, Rutu; Gupta, Ishita; Khandait, Harshwardhan; Yadav, Ashima; Yadav, Anmol (2020): COVID-19 and its Modes of Transmission. In: SN comprehensive clinical medicine, pp. 1–4. DOI: 10.1007/s42399-020-00498-4.
- [5] Cirrincione, Luigi; Plescia, Fulvio; Ledda, Caterina; Rapisarda, Venerando; Martorana, Daniela; Moldovan, Raluca Emilia et al. (2020) COVID-19 Pandemic. Prevention and Protection Measures to Be Adopted at the Workplace. In : Sustainability, vol. 12, n° 9, p. 3603. DOI: 10.3390/su12093603.
- [6] World Health 1211 Geneva 27: COVID-19 STRATEGYUPDATE.Switzerland,WHO in Emergencies:[www.who.int/emergencies/en](http://www.who.int/emergencies/en).
- [7] Lepelletier, Didier; Grandbastien, Bruno; Romano-Bertrand, Sara; Aho, Serge; Chidiac, Christian; Géhanno, Jean-François; Chauvin, Franck (2020): What face mask for what use in the context of COVID-19 pandemic? The French guidelines. In: The Journal of hospital infection. DOI: 10.1016/j.jhin.2020.04.036.
- [8] Zheng Jiang, Menghan Hu, Lei Fan, Yaling Pan, Wei Tang, Guangtao Zhai, Yong Lu:Combining Visible Light and Infrared Imaging for Efficient Detection of Respiratory Infections such as COVID-19 on Portable Device. CoRR abs/2004.06912 (2020).
- [9] Michael J. Horry, Subrata Chakraborty, Manoranjan Paul, Anwaar Ulhaq, Biswajeet Pradhan, Manas Saha, Nagesh Shukla :COVID-19 Detection Through Transfer Learning Using Multimodal Imaging Data. IEEE Access 8: 149808-149824 (2020).
- [10] Laguarda, Jordi; Hueto, Ferran; Subirana, Brian (2020): COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings. In: IEEE Open J. Eng. Med. Biol., 1, pp. 275–281. DOI: 10.1109/OJEMB.2020.3026928.
- [11] Loey, Mohamed; Manogaran, Gunasekaran; Taha, Mohamed Hamed N.; Khalifa, Nour Eldeen M. (2021): Fighting against COVID-19. A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection. In: Sustainable cities and society, 65, p. 102600. DOI: 10.1016/j.scs.2020.102600.
- [12] NenadPetrović and ĐorđeKocić: IoT-based System for COVID-19 Indoor SafetyMonitoring . Conference: IcETRAN 2020.
- [13] RidiArif, KoekoehSantoso andDhani S. Wibawa: Rats Development of Contactless Thermal Detector for Animal: Comparison of Three Sensor Types. (ICVAES 2020).
- [14] Dong, Ke; Zhou, Chengjie; Ruan, Yihan; Li, Yuzhi (2020): MobileNetV2 Model for Image Classification. In: 2020 2nd International Conference on Information Technology and Computer Application (ITCA):IEEE, pp. 476–480.
- [15] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, Liang-Chieh Chen: MobileNetV2: Inverted Residuals and Linear Bottlenecks. CVPR 2018: 4510-4520.
- [16] Sidra Mehtab and Jaydip Sen : Face Detection Using OpenCV and Haar Cascades Classifiers. March 2020 DOI: 10.13140/RG.2.2.26708.83840
- [17] Anam, K. (2020). Smart Home Pengendali Lampu Rumah Berbasis SMS Gateway dan Arduino Menggunakan Smartphone Android. *Jurnal Ilmiah Informatika*, 5(2), 122-132. <https://doi.org/10.35316/jimi.v5i2.945>.
- [18] Lux, Mathias; Bertini, Marco (2019): Open source column: deep learning with Keras. In: SIGMultimedia Rec. n. 4, 10, p. 7. DOI: 10.1145/3310195.3310202.

- [19] Face-Mask-Detection:<https://github.com/balajisrinivas/Face-Mask-Detection/tree/master/dataset>.
- [20] Xie, Yuanlun; He, Majun; Ma, Tingsong; Tian, Wenhong (2021) Optimal distributed parallel algorithms for deep learning framework Tensorflow. In : Applied Intelligence, vol. 521, n° 7553, p. 436. DOI: 10.1007/s10489-021-02588-9.
- [21] Liu, Mingliang; Grana, Dario (2019) Accelerating geostatistical seismic inversion using TensorFlow. A heterogeneous distributed deep learning framework. In : Computers & Geosciences, vol. 124, n° 6, p. 37–45. DOI: 10.1016/j.cageo.2018.12.007.
- [22] M, Hossin; M.N, Sulaiman (2015) A Review on Evaluation Metrics for Data Classification Evaluations. In: International Journal of Data Mining & Knowledge Management Process, vol. 5, n° 2, p. 1–11. DOI: 10.5121/ijdkp.2015.5201.
- [23] Powers, David Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation.
- [24] Muhamad Yani, Budhi Irawan, S, Si., M.T and Casi Setiningsih, S.T., M.T :Application of Transfer Learning Using Convolutional Neural Network Method for Early Detection of Terry's Nail(2019). DOI:10.1088/1742-6596/1201/1/012052.
- [25] Militante, Sammy V.; Dionisio, Nanette V. (2020): Real-Time Facemask Recognition with Alarm System using Deep Learning. In: 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC). 2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC). Shah Alam, Malaysia, 08/08/2020 - 08/08/2020: IEEE, pp. 106–110.
- [26] Cmak Zeelan Basha, B. N. Lakshmi Pravallika, E. Bharani Shankar:An Efficient Face Mask Detector with PyTorch and Deep Learning. EAI Endorsed Trans. Pervasive Health Technol. 7(25): e4 (2021).
- [27] Sudianto, Agus; Jamaludin, Zamberi; Abdul Rahman, Azrul Azwan; Novianto, Sentot; Muharrom, Fajar (2020) Smart Temperature Measurement System for Milling Process Application Based on MLX90614 Infrared Thermometer Sensor with Arduino. In : Journal of Advanced Research in Applied Mechanics, vol. 72, n° 1, p. 10–24. DOI: 10.37934/aram.72.1.1024.
- [28] Karzan A. Raza; Wrya Monnet: Moving objects detection and direction-finding with HC-SR04 ultrasonic linear array.(IEC2019).
- [29] Cryptool:  
<https://play.google.com/store/apps/details?id=io.github.nfdz.cryptool&hl=fr&gl=US>.
- [30] Tina Chen :Reducing COVID-19 Transmission Through Cleaning and DisinfectingHousehold SurfacesFinal. Oct 14 2020.

# Human Emotion Recognition by Integrating Facial and Speech Features: An Implementation of Multimodal Framework using CNN

P V V S Srinivas, Pragnyaban Mishra

Department of Computer Science and Engineering  
Koneru Lakshmaiah Education Foundation (KLEF)  
Guntur, India

**Abstract**—This Emotion recognition plays a prominent role in today's intelligent system applications. Human computer interface, health care, law, and entertainment are a few of the applications where emotion recognition is used. Humans convey their emotions in the form of text, voice, and facial expressions, thus developing a multimodal emotional recognition system playing a crucial role in human-computer or intelligent system communication. The majority of established emotional recognition algorithms only identify emotions in unique data, such as text, audio, or image data. A multimodal system uses information from a variety of sources and fuses the information by using fusion techniques and categories to improve recognition accuracy. In this paper, a multimodal system to recognise emotions was presented that fuses the features from information obtained from heterogenous modalities like audio and video. For audio feature extraction energy, zero crossing rate and Mel-Frequency Cepstral Coefficients (MFCC) techniques are considered. Of these, MFCC produced promising results. For video feature extraction, first the videos are converted to frames and stored in a linear scale space by using a spatial temporal Gaussian Kernel. The features from the images are further extracted by applying a Gaussian weighted function to the second momentum matrix of linear scale space data. The Marginal Fisher Analysis (MFA) fusion method is used to fuse both the audio and video features, and the resulted features are given to the FERCNN model for evaluation. For experimentation, the RAVDESS and CREMAD datasets, which contain audio and video data, are used. Accuracy levels of 95.56, 96.28, and 95.07 on the RAVDESS dataset and accuracies of 80.50, 97.88, and 69.66 on the CREMAD dataset in audio, video, and multimodal modalities are achieved, whose performance is better than the existing multimodal systems.

**Keywords**—Emotion recognition; multimodal; fusion; MFCC; MFA; FERCNN; CREMAD; RAVDESS

## I. INTRODUCTION

Emotion recognition is the process of determining a person's emotional state. Affective computing and human-computer interaction (HCI) applications rely heavily on it [1]. In recent studies, emotion identification has sparked increased attention in academics and the commercial sector [2]. It is used in a variety of applications, including analysis of Twitter, tutoring systems, playing video games, prediction of consumer satisfaction, and military healthcare [3-5].

Speech or audio emotion recognition has been employed in medical studies to examine the changes in the emotions of depressed patients and in children who are having communication difficulties. It can also be used to warn the drivers during driving when the condition of the driver is fatigued to avoid accidents. Low level information from the speech or audio signals is extracted by the speech or audio emotion recognition system to comprehend the emotion status. Compilation of databases related to emotions, extraction of emotional features from the speech or audio signal, reduction of features by using dimensionality reduction techniques, and classification of emotions into respective classes are all part of this classification problem based on speech or audio signal sequences. K nearest neighbour, Gaussian mixture model, Support vector machines, and artificial neural networks are some of the traditional techniques that are used for speech or audio emotional recognition and are not that efficient because human emotions have high complexity and uncertainty [6].

About 93% of communication with humans is done through nonverbal means such as voice tone, facial expressions, and body language [7]. Identifying emotions through facial expressions which has been extensively studied [8][9] resulted in higher accuracies by making the changes at the pre-processing stage. To reduce overfitting during the training stage, adding dropout to the CNN model plays a prominent role in reducing overfitting during training [10]. Extracting of faces from the chain of video sequences and extracting the features from the resulted images are the steps followed in general to detect the emotions of the faces in the video sequences [11]. The robust face detection algorithm [12], the AdaBoost learning algorithm [13], and the spatial template tracker [14] are some of the techniques used in detecting the faces in the video. Fisher vectors, Active Shape model, Active Appearance model, local binary patterns, principal component analysis [15] and Gaussian mixture model [16] are some of the methods that are used for feature extraction in facial images. Occlusions and light changes may also lead the identification technique to be misled. If the emotion is to be identified through speech, ambient noise and differences in the voices of different participants are major factors that might affect the final recognition result. According to both physiological and psychological research, humans need both audio and visual signals to correctly understand emotions for which multimodal systems that fuse audio and video signals can be used.

Thanks to recent research interest in multimodal systems, the limitations of monomodal systems [17] [18] have been overcome. The information obtained from different modalities at different levels of fusion was fused by multimodal systems. The different fusion levels are classified into two different categories, namely: matching prior to fusion and matching after fusion. Feature level and sensor level fusion techniques [19] come under the first category, and decision, rank, and score level fusion techniques come under the second category. To combine the audio and video features of the multimodal, a fusion method that takes advantage of both decision and feature-level fusion was developed. Latent space fusion methods preserve analytical or numerical correlation between the different modalities and store them in a common latent space.

## II. RELATED STUDIES AND MOTIVATIONS

Many attempts have been made by researchers to enhance emotion identification using a combination of audio and visual information [20]. According to [21], audio-visual emotion detection may be categorized as kernel-based, feature level, model-level, decision-level, score-level, and hybrid level fusion techniques. In this paper, we focus on latent-space fusion methods and multimodal recognition to detect emotions. Multimodal emotion recognition systems consistently outperform unimodal systems [22], [23], and [24]. Although there are certain benefits to using multimodal affective systems, they also face some important challenges [24]. Selecting the modalities that result in the best combinations is the area that has been focused on in recent studies [25]. CREMA-D [26], RAVDESS [27], and SAVEE [28] are some of the existing multimodal datasets that have been considered for research in recent times. A multimodal method by Cid et al. [29] used tempo, pitch, and energy feature extraction techniques to extract the audio features and a Bayesian classification method to classify the emotions. Edge-based characteristics are obtained from visual images to classify them in the SAVEE database.

Gharavian et al. [30] evaluated the performance of a neural network called FAMNN. MFCC, Zero Crossing Rate, and pitch are some of the audio feature extraction techniques used to extract the audio features. Visual information is obtained by using marker positions on the face concept, and the resulted features are given to a feature selection algorithm (FCBF). For audio features, Huang et al. [31] used prosodic and frequency domains, while for facial expression description, they used geometry and appearance-based features. Using a back-propagation neural network, each feature vector was utilised to train a single-modal classifier. They suggested a genetic learning-based collaborative decision-making model, which was compared to concatenated equal weighted choice fusion, BPN learning-based weighted decision fusion, and feature fusion methods. The audio spectrum features are obtained from BERT and CNN and are combined in parallel to form a multimodal [32].

A HGFM method was proposed by Xu [33], which fuses the hand-crafted features and the features extracted from the gated recurrent unit. The key frame videos are summarized by

the method proposed by Noroozi [34] which uses a CNN model and the concept of stack fashion or late fusion for detecting the emotions. Xu et al. [35] proposed a multi-hop memorized network that describes the single-modality and cross-modality interactions among the three different feature domains in aspect-level sentimental analysis of a multimodal system. Zadeh et al. [36] introduced a tensor fusion network that uses the product of audio, visual and image elements to represent multimodal fusion information.

RMFN, a multistage recurrent network for fusion described by Liang et al. [37], divides the multimodal fusion into various stages that utilize LSTM to record multimodal interactions in both synchronous and asynchronous modes. Liu et al. [38] lowered the computational complexity of the parameters by using a low-rank multimodal fusion approach that employs a low-rank tensor to relieve the increased computational cost of considering all three modalities. Poria et al. [39] used LSTM to isolate audio, video and text elements before combining them in a multi-level architecture. Ghosal et al. [40] developed a multi-attention recurrent network architecture for multimodal representation that learns features through attention. Tsai et al. [41] suggested learning interactions between modalities by employing multimodal transformers to construct an attention-based cross-modal architecture.

By using the RAVDESS dataset Fu Z et al. [47], R. Chatterjee et al. [48], Chang X et al. [49], Wang W et al. [50] achieved test accuracies of 75.76, 90.48, 91.4, and 89.8 on their respective multimodal systems. Ghaleb E et al. [52], He G et al. [53] proposed multimodal systems which resulted in test accuracies of 66.5 and 64 on the CREMAD dataset. Rory Beard et al. [51] proposed a multimodal where CREMAD and RAVDSR datasets are used for experimentation and resulted in test accuracies of 65.0 and 58.3, respectively.

## III. RESEARCH METHOD

### A. Dataset Description

CREMAD and RAVDESS datasets are used for experimentation and evaluation purposes. Both datasets consist of data related to the emotions of actors in both audio and video modes. Angry, disgust, fear, happy, neutral, and sad are the common emotions present in both datasets in both modes, whereas RAVDESS audio data consists of two more emotions, calm, and surprise. CREMAD consists of 22326 and 60359 emotions related to audio and video. RAVDESS consists of 4321 and 45225 emotions related to audio and video. A detailed overview of the datasets is given in Table I below.

### B. Image Feature Extraction

From the given set of video sequences of the multimodal dataset the videos should be converted into images and then facial features should be extracted from the images. The detailed description of the features is extracted from the videos is given below.

From the given set of facial emotion videos  $f_{vid}$  of a multimodal dataset, the images are represented in linear scale space  $L_{ss}$  which is obtained by convoluting  $f_{vid}$  with 3 dimensional Gaussian Kernel.

TABLE I. DESCRIPTION OF CREMAD AND RAVDESS DATASETS

| Name of The Dataset | Emotion Type | Data mode and Number of Emotions |                  |
|---------------------|--------------|----------------------------------|------------------|
|                     |              | Audio Mode                       | Video/Image Mode |
| CREMAD Dataset      | Angry        | 3510                             | 10472            |
|                     | Disgust      | 4116                             | 10098            |
|                     | Fear         | 3918                             | 10626            |
|                     | Happy        | 3709                             | 9661             |
|                     | Neutral      | 3666                             | 10867            |
|                     | Sad          | 3417                             | 8635             |
| RAVDESS Dataset     | Angry        | 476                              | 7603             |
|                     | Calm         | 524                              | NA               |
|                     | Disgust      | 628                              | 7885             |
|                     | Fear         | 542                              | 7394             |
|                     | Happy        | 610                              | 7784             |
|                     | Neutral      | 385                              | 7419             |
|                     | Sad          | 559                              | 7140             |
|                     | Surprise     | 596                              | NA               |

$$L_{ss}(\cdot; \sigma_{L_{ss}}^2, \tau_{L_{ss}}^2) = \text{Gau}_k(\cdot; \sigma_{L_{ss}}^2, \tau_{L_{ss}}^2) * f_{\text{vid}}(\cdot) \quad (1)$$

linear scale space,  $f_{\text{vid}}$  is video sequence,  $\sigma_{L_{ss}}^2$  is Spatial variance,  $\tau_{L_{ss}}^2$  is Temporal variance,  $\text{Gau}_k$  is Spatial Temporal Gaussian Kernel.

$$\text{Gau}_k(x, y, t_d; \sigma_{L_{ss}}^2, \tau_{L_{ss}}^2) = \exp(-(x^2 + y^2)/2\sigma_{L_{ss}}^2 - t_d^2 / 2 \tau_{L_{ss}}^2) \quad (2)$$

Whereas  $x$  and  $y$  represents the axis of the frames that are obtained from the facial input video sequence  $f_{\text{vid}}$ ,  $t_d$  denotes the axis if time in the temporal domain

A method proposed by Forstner and Harris [42] [43] considers a Gaussian window to identify distinct points of the image which in turn determines the locations in  $f_{\text{vid}}$  when there are significant changes in the intensity of image in the given space and time domains when sliding the Gaussian window in various directions. The distinct points can be detected by convoluting Spatial-Temporal Second Momentum matrix with the given Gaussian weighted function  $\text{Gau}_k(\cdot; \sigma_i^2, \tau_i^2)$ .

The Spatial-Temporal Second Momentum matrix is  $3 \times 3$  dimensional matrix and is given as

$$\begin{bmatrix} L_{ssx}^2 & L_{ssx}L_{ssy} & L_{ssx}L_{sst} \\ L_{ssx}L_{ssy} & L_{ssy}^2 & L_{ssy}L_{sst} \\ L_{ssx}L_{sst} & L_{ssy}L_{sst} & L_{ssz}^2 \end{bmatrix} \quad (3)$$

And the distinct points identification is given by

$$\mu_{\text{ch}} = \text{Gau}_k(\cdot, \sigma_i^2, \tau_i^2) * \left( \begin{bmatrix} L_{ssx}^2 & L_{ssx}L_{ssy} & L_{ssx}L_{ssz} \\ L_{ssx}L_{ssy} & L_{ssy}^2 & L_{ssy}L_{ssz} \\ L_{ssx}L_{ssz} & L_{ssy}L_{ssz} & L_{ssz}^2 \end{bmatrix} \right) \quad (4)$$

Where  $L_{ssx}$ ,  $L_{ssy}$  &  $L_{ssz}$  are first order derivatives that are defined as follows

$$L_{ssx}(\cdot, \sigma_{L_{ss}}^2, \tau_{L_{ss}}^2) = \partial_x(\text{Gau}_k * f_{\text{vid}}) \quad (5)$$

$$L_{ssy}(\cdot, \sigma_{L_{ss}}^2, \tau_{L_{ss}}^2) = \partial_y(\text{Gau}_k * f_{\text{vid}}) \quad (6)$$

$$L_{ssz}(\cdot, \sigma_{L_{ss}}^2, \tau_{L_{ss}}^2) = \partial_z(\text{Gau}_k * f_{\text{vid}}) \quad (7)$$

Where  $\sigma_i^2 = S_{\text{ssk}} * \sigma_{L_{ss}}^2$ ,  $\tau_i^2 = S_{\text{ssk}} * \tau_{L_{ss}}^2$  and  $S_{\text{ssk}}$  is a constant

The existence of distinct points in the  $f_{\text{vid}}$  is indicated by the eigen values  $\lambda_1, \lambda_2, \lambda_3$  that can hold larger values. In the Spatial-Temporal domain the variations that are existing in the intensity of image are obtained by concatenating the  $\text{trace}_{L_{ss}}$  and determinant of  $\mu_{\text{ch}}$  which is given as

$$H_{fn} = |(\mu_{\text{ch}})| - K * \text{trace}_{L_{ss}}^3(\mu_{\text{ch}}) = \lambda_1 * \lambda_2 * \lambda_3 - k(\lambda_1 + \lambda_2 + \lambda_3) \quad (8)$$

$K$  is a constant and the function  $H_{fn}$  is normalized such that the effect of variations in the images due to illumination can be removed

### C. Audio Feature Extraction

Zero crossing rate (ZCR), Mel Frequency Spectrum Coefficient (MFCC), pitch and energy are some of the feature extraction techniques used to extract the features of the emotions from the given audio signal.

1) *Zero crossing rate*: The number of times the audio signal crosses the zero-line, x-axis, is referred to as the zero-crossing rate, and it is stated as follows.

$$Z_{t_n} = \frac{1}{2N} * \sum_{n=1}^N \left| \text{Sign}_{\text{Aud}}(x_{\text{Audt}}(n)) - \text{Sign}_{\text{Aud}}(x_{\text{Audt}}(n-1)) \right| \quad (9)$$

$$\text{Sign}_{\text{Aud}}(x_{\text{Audt}}) = \begin{cases} 1 & \text{if } x_{\text{Audt}} > 0 \\ 0 & \text{Otherwise} \end{cases} \quad (10)$$

Where,  $t_n \in [t_{n1}, t_{n2}]$ ,  $x_{\text{Audt}(t_n)}$  is the respective audio signal that was divided into segments by using a sliding window that was having a length of  $T$ ,  $n \in [0, N]$  and  $x_{\text{Audt}}(n)$  is the  $t_n^{\text{th}}$  Segments time sequence

2) *MFCC (Mel frequency cestrum coefficient)*: The coefficients of the corresponding spectral form of the audio stream are represented using a nonlinear Mel scale. The Mel frequency was used to analyse cepstral coefficients, and the steps below were followed.

*Step 1: Audio Signals are splitted into frames by using fixed shift and window sizes.*

*Step 2: Fast Fourier Transform (FFT)*

*for each frame is calculated.*

*Step 3: Frequencies are based on the Mel Scale used.*

*Step 4: Logarithm of the resulted output of Step 3 is calculated.*

*Step 5: Discrete Cosine Transform (DCT) for each frame is calculated.*

Acoustic tube characteristics pitch and energy are exhibited by MFCC that contains great amount of emotional information which plays a key role in emotion recognition.

3) *Pitch*: It depicts the signal's fundamental frequency [44]. The valence of an audio stream is connected to its rhythm and average pitch from an emotional standpoint. For example, higher amount of pitch may be associated to discomfort, lower standard deviation to sadness and usually happiness and discomfort are having higher talk and pitch rates whereas sadness can be represented by lower talk and pitch rates [45]. Autocorrelation is used to calculate the pitch of the audio signal and is given as follows.

$$x_{\text{Aud}}[n] \text{ be Stochastic Process Sinusoidal function}$$

$$x_{\text{Aud}}[n] = \text{Cos}(w_0 n + \emptyset) \text{ and the autocorrelation of } x_{\text{Aud}}[n] \text{ is given as}$$

$$R_{\text{Aud}}[t] = E\{x_{\text{Aud}}^*[n] * x_{\text{Aud}}^*[n+t]\} \quad (11)$$

$$= \frac{1}{2} \text{cos}(w_0 t)$$

Maximum of the autocorrelation value is used to calculate the pitch,  $S_{\text{Aud}}$  Samples are used to calculate the estimate of  $R_{\text{Aud}}[t]$

$$R_{\text{Aud}}^{\wedge}[t] = \frac{1}{S_{\text{Aud}}} * \sum_{S_{\text{Aud}}=0}^{S_{\text{Aud}}-|t|} (W_{\text{Aud}}[S_{\text{Aud}}] * x_{S_{\text{Aud}}} * W_{\text{Aud}}[S_{\text{Aud}} + |t|]) \quad (12)$$

$W_{\text{Aud}}[S_{\text{Aud}}]$  is window length of  $S_{\text{Aud}}$  the Expected value of

$$R_{\text{Aud}}^{\wedge}[t] \text{ is given as}$$

$$E_{\text{Aud}}\{R_{\text{Aud}}^{\wedge}[t]\} = \left(1 - \frac{|t|}{S_{\text{Aud}}}\right) * \frac{\text{Cos}(w_{\text{Aud}0} * S_{\text{Aud}})}{2}, |t| < S_{\text{Aud}} \quad (13)$$

4) *Energy*: It represents the signal's intensity or total energy. From an emotional standpoint, an audio signal having exciting emotions (e.g., pain or happiness) has more energy than an audio signal containing sadness or fatigued feelings [46]. The energy of the audio signal  $x_{\text{Aud}t}(n)$  is given as

$$\text{Energy}_{\text{Aud}} = \sqrt{\frac{1}{N} * \sum_{n=1}^N (x_{\text{Aud}t}(n)^2)} \quad (14)$$

#### D. Feature Level Fusion

From the features obtained from audio and video signals, only a few portions of the features are related to emotions. Personality, age, gender, and many other features are obtained from audio and video signals, which may impact the quality of recognition of the emotions that are used in the model for training. Feature Level Latent Space methods are one of the existing categories of methods that are used to find the common features related to emotions and map them into the required latent space. By maximizing the cross correlation of the respective features and by minimizing the feature distance or by taking the normalization of the features, they can be used in feature level fusion. Marginal Fisher Analysis (MFA) is a supervised method that is used for audio video feature level for fusion by extracting the required features from the respective modalities. The process of  $MFA^s$  feature level fusion is given as below.

Information related to class labels is used in latent space generation. The compactness in the intra class is given as

$$S_{\text{compact}} = \sum_i \sum_{i \in N_{k_1}^+} \|W_{\text{AV}}^T x_i - W_{\text{AV}}^T x_i\|^2$$

$$= 2 W_{\text{AV}}^T X_{\text{AV}} (D^{\text{AV}} - S^{\text{AV}}) * X_{\text{AV}}^T W_{\text{AV}} \quad (15)$$

$X_{\text{AV}} = \{x_1, x_2, \dots, x_n\}$  is the frame set,  $N$  is the total samples and  $N_{k_1}^+$  is  $k_1$  in the same class.

$$S_{ij}^{\text{AV}} = \begin{cases} 1 & \text{if } i \in N_{k_1}^+(j) \\ 0 & \text{Otherwise} \end{cases} \quad (16)$$

$$D_{ij}^{\text{AV}} = \sum_j S_{ij}^{\text{AV}} \quad (17)$$

And the Inter-Class Separability is given by

$$I_{\text{CP}} = \sum_i \sum_{(i,j) \in P_{k_2}(c_i)} \|W_{\text{AV}}^T x_i - W_{\text{AV}}^T x_j\|^2 \quad (18)$$

$$= 2 W_{\text{AV}}^T X_{\text{AV}} (D_{\text{AV}}^{\text{P}} - S_{\text{AV}}^{\text{P}}) * W_{\text{AV}}^T W_{\text{AV}}$$

$c_i$  is the emotion of class  $i$ ,  $P_{k_2}(c_i)$  is the set of  $K_2$  nearest pairs and  $S_{\text{AV}}$  is given by

$$S_{\text{AV}ij}^{\text{P}} = \begin{cases} 1 & \text{if } (i,j) \in P_{k_2}(c_i) \\ 0 & \text{Otherwise} \end{cases} \quad (19)$$

And the objective function is given as follows

$$W_{\text{AV}}^{\wedge} = \arg_{W_{\text{AV}}} \left\{ \min \left\{ \frac{W_{\text{AV}}^T X_{\text{AV}} (D^{\text{AV}} - S^{\text{AV}}) X_{\text{AV}}^T W_{\text{AV}}}{W_{\text{AV}}^T X_{\text{AV}} (D_{\text{AV}}^{\text{P}} - S_{\text{AV}}^{\text{P}}) X_{\text{AV}}^T W_{\text{AV}}} \right\} \right\} \quad (20)$$

And the optimal solution is given by

$$Y_{\text{AV}} = X_{\text{AV}}^T W_{\text{AV}} \quad (21)$$

$$L_{\text{AV}} \cdot Y_{\text{AV}} = \lambda L_{\text{AV}}^{\text{P}} \cdot \quad (22)$$

Where  $L_{\text{AV}} = D^{\text{AV}} - S^{\text{AV}}$  and  $L_{\text{AV}}^{\text{P}} = D_{\text{AV}}^{\text{P}} - S_{\text{AV}}^{\text{P}}$  are called Laplacian matrices for  $W_{\text{AV}}$  and  $W_{\text{AV}}^{\text{P}}$

#### E. Proposed CNN Architecture

The proposed CNN architecture consists of four fully connected layers, one flattening layer and two dense layers. All the fully connected layers are interconnected with each other where the output features obtained from each fully connected layer are given as an input to the next fully connected layer. The inputs to the first fully connected layer are audio, video, and multimodal features that are obtained during pre-processing by applying the audio feature, image feature, and feature level fusion extraction techniques described in the above sections. The first fully connected layer consists of convolution and max polling layers, and the representation of the first fully connected layer is given as

$$\text{Out}_{\text{conv}1} = \text{Act}(\sum_i L_{\text{AV}} * W_{ij}^n) \quad (23)$$

Where  $\text{Out}_{\text{conv}1}$  is the output of the convolutional layer,  $\text{Act}$  is the activation function,  $L_{\text{AV}}$  is the latent space or latent features obtained after applying feature level fusion, and  $W_{ij}^n$  is the set of weights associated with the convolutional layer

$$\text{Out}_{\text{conv}f1} = \text{Max polling}\{\text{Out}_{\text{conv}1}\} \quad (24)$$

$Out_{convf1}$  is the output obtained from the max polling layer, where the input is  $Out_{conv1}$ , the first convolutional layer output. The output of the first fully connected layer  $Out_{Maxpoll1}$  is given as input to the second fully connected layer, which consists of convolutional, max polling, and dropout layers, and the representation of the second fully connected layer is given as

$$Out_{conv2} = Act(\sum_i Out_{Maxpoll1} * W_{ij}^{2n}) \quad (25)$$

$$Out_{Maxpoll2} = \text{Max polling}\{Out_{conv2}\} \quad (26)$$

$$Out_{conv2f} = Act((Out_{Maxpoll2} * Drop(0.2))) * W^{[2n+1]} \quad (27)$$

$Out_{conv2f}$  is the output of the second fully connected layer,  $Drop(0.2)$  means that 20% of the features are dropped from the output of the max polling layer, and  $W^{[2n+1]}$  are associated weights used.

$Out_{conv2f}$  the output of second fully connected layer, is given as input to the third fully connected layer which consists of the same layers as second fully connected layer and the output of the third fully connected layer is given as

$$Out_{conv3f} = Act((Out_{Maxpoll3} * Drop(0.2))) * W^{[2n+2]} \quad (28)$$

$Out_{conv3f}$  is given as input to the fourth fully connected layer which consists of a convolution and max polling layers and the output is given as

$$Out_{conv4} = Act(\sum_i Out_{conv3f} * W_{ij}^n) \quad (29)$$

$$Out_{conv4f} = \text{Max polling}\{Out_{conv4}\} \quad (30)$$

The output of the fourth fully connected layer is flattened by giving to a flatten layer and the output is represented as

$$\text{Flatten}_{CNN} = \text{Flatten}(a_1 Out_{convf1}, a_2 Out_{convf2}, a_3 Out_{convf3}, a_4 Out_{convf4}) \quad (31)$$

The output of a flattening layer is given to a dense layer and a dropout of 20% is applied to the output obtained from the dense layer. The resultant features are given as input to the next dense layer where the output is classified. Relu activation function is used in the dense layers that are used in between, and a SoftMax activation function is used in the final dense output layer. The representation of the dense, dropout, and final output layers is as follows:

$$Out_{Dense1}^1 = \text{Dense}(Den_N, Act_{Relu}(\text{Flatten}_{CNN})) \quad (32)$$

$$Out_{Drop}^1 = Act(Out_{Dense1}^1 * Drop(0.2)) * W \quad (33)$$

$$Out_F^1 = \text{Dense}(Den_C, Act_{Softmax}(Out_{Drop}^1)) \quad (34)$$

$Out_{Dense1}^1$  is the output of the dense layer,  $Out_{Drop}^1$  is the output of dropout layer  $Out_F^1$  is the final classified output. The architecture of the proposed CNN is given in the Fig. 1.

```

Model: "sequential"

Layer (type) Output Shape Param #

conv1d (Conv1D) (None, 168, 256) 1824

max_pooling1d (MaxPooling1D) (None, 53, 256) 0

conv1d_1 (Conv1D) (None, 51, 128) 98432

max_pooling1d_1 (MaxPooling1 (None, 17, 128) 0

dropout (Dropout) (None, 17, 128) 0

conv1d_2 (Conv1D) (None, 15, 64) 24648

max_pooling1d_2 (MaxPooling1 (None, 5, 64) 0

dropout_1 (Dropout) (None, 5, 64) 0

conv1d_3 (Conv1D) (None, 3, 64) 12352

max_pooling1d_3 (MaxPooling1 (None, 3, 64) 0

conv1d_4 (Conv1D) (None, 1, 32) 6176

max_pooling1d_4 (MaxPooling1 (None, 1, 32) 0

flatten (Flatten) (None, 32) 0

dense (Dense) (None, 256) 8448

dropout_2 (Dropout) (None, 256) 0

dense_1 (Dense) (None, 6) 1542

Total params: 152,614
Trainable params: 152,614

```

Fig. 1. Proposed CNN Architecture.

### F. Data Preprocessing

For experimentation, the RAVDESS and CREMAD datasets are used in this paper. The datasets contain data related to audio and video emotions of various actors, and the description of the data is given in the dataset description section of the same module. The features of the video and audio data are obtained by using the image feature extraction and audio feature extraction methods explained above. There is dissimilarity in the number of features obtained from audio and video datasets. There are more features in the resultant dataset of video images when compared to audio files. A dimensionality reduction technique is applied to the image set to reduce the number of features so that the same number of features is present in the audio and video resultant datasets. Finally, a multimodal dataset is obtained by combining the resultant features of audio and video from the respective datasets by using the feature-level fusion technique that was explained in Section D, namely the "Feature Level Fusion" of the same module. The features obtained after applying Feature Level Fusion are given to the proposed CNN Model for Evaluation, and the description of the proposed CNN Model is explained in Section E, named "Proposed CNN Architecture." Fig. 2 gives the workflow of the proposed work done in this paper.

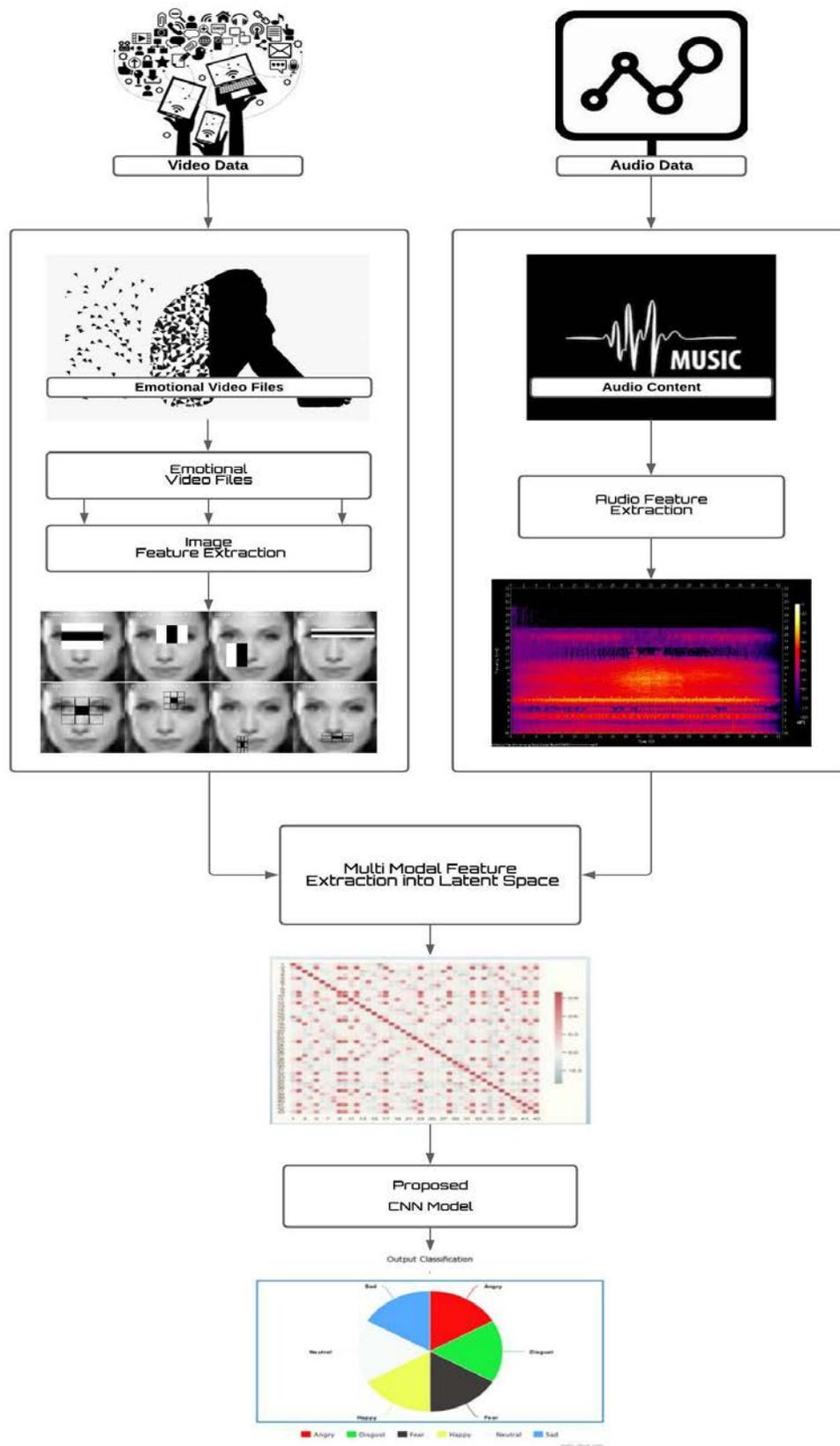


Fig. 2. Workflow of the Proposed Method.

IV. EXPERIMENTATION AND RESULTS

Fig. 3(a) and (b), 3(c) and (d) and 3(e) and (f) represent training and testing accuracy and loss comparisons in audio, video, and multimodal modes on the RAVDESS dataset. Test accuracies of 95.96, 96.28, and 95.07 were observed. On the CREMAD dataset, train and test accuracies and train and test accuracies losses are shown in Fig. 4(a) and (b), 4(c) and (d), and 4(e) and (f) represent training and testing accuracy and loss comparisons in audio, video, and multimodal modes. Test accuracies of 80.70, 97.88, and 69.66 were observed. A detailed description of the results is given in Table II.

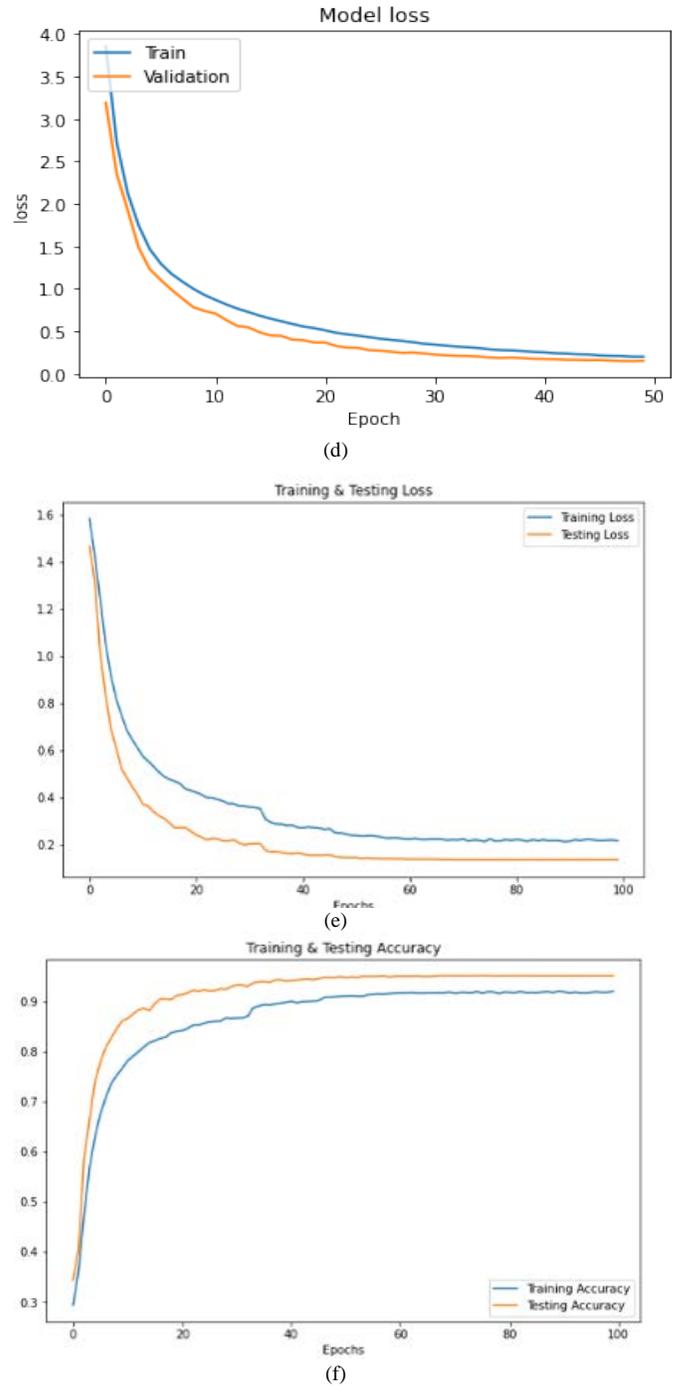
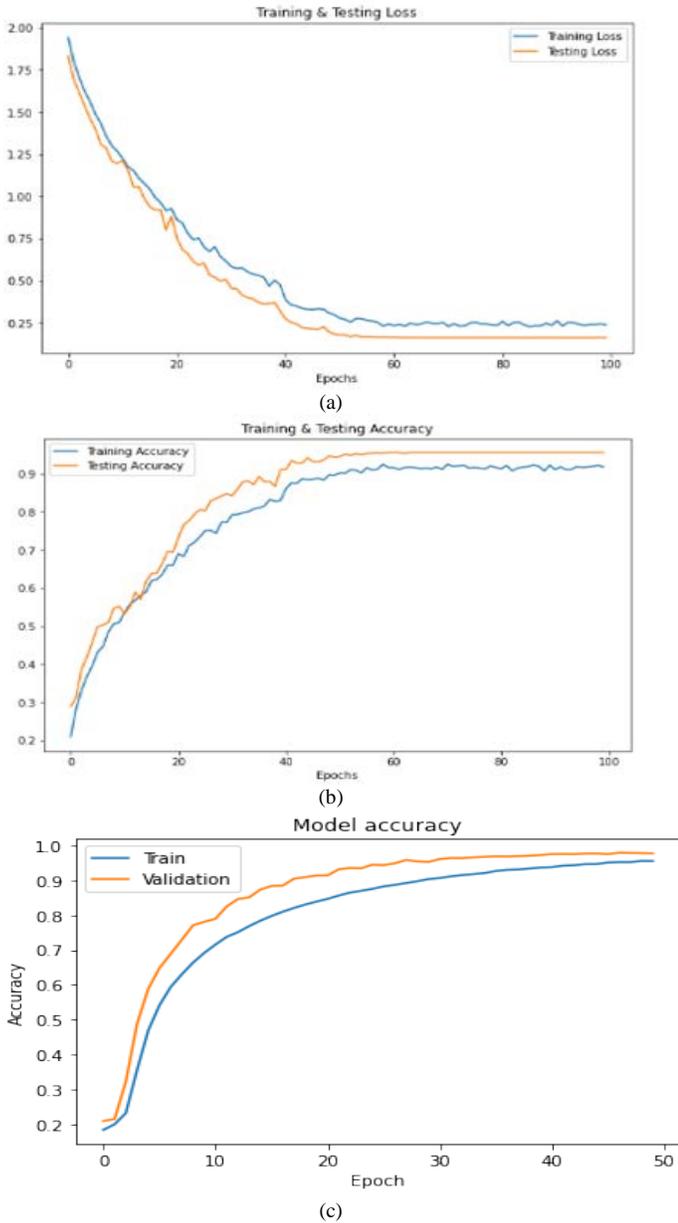


Fig. 3. (a) Training and Testing Loss of Audio Data in RAVDESS Dataset, (b) Training and Testing Accuracy of Audio Data in RAVDESS Dataset, (c) Training and Testing Accuracy of Video Data in RAVDESS Dataset, (d) Training and Testing Loss of Video Data in RAVDESS Dataset, (e) Training and Testing Loss of Multi Modal Data in RAVDESS Dataset, (f) Training and Testing Accuracy of Multi Modal Data on.

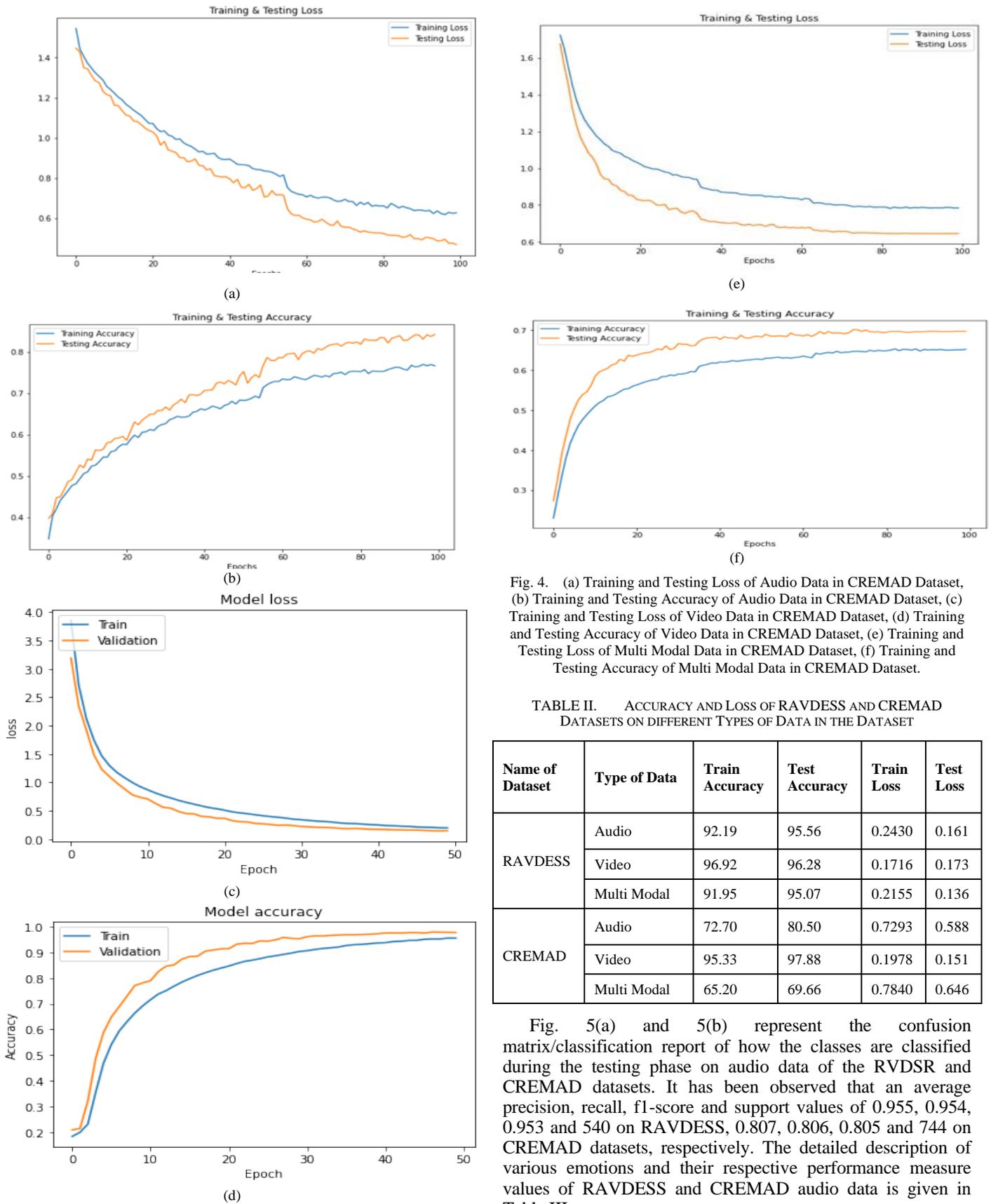
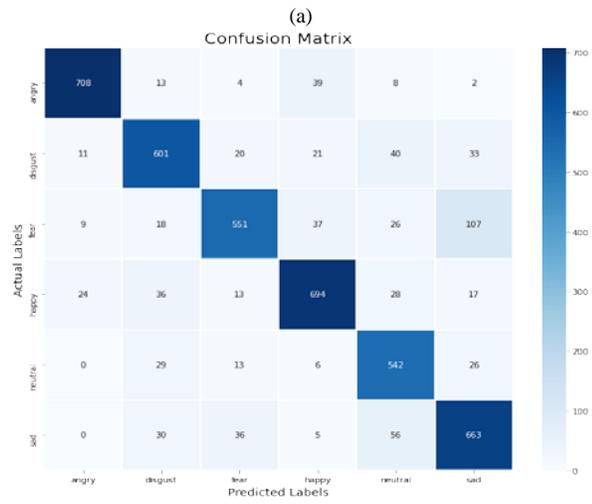
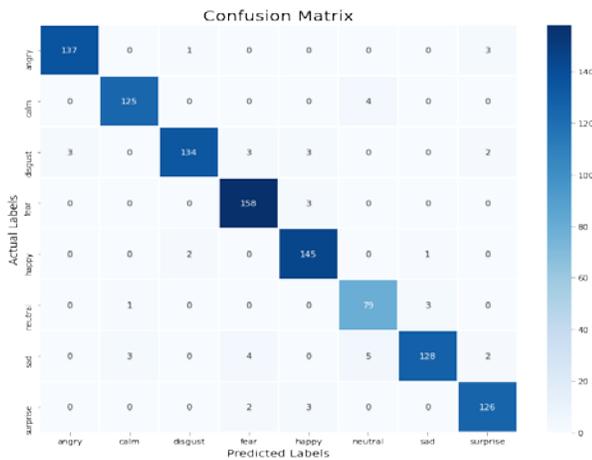


Fig. 4. (a) Training and Testing Loss of Audio Data in CREMAD Dataset, (b) Training and Testing Accuracy of Audio Data in CREMAD Dataset, (c) Training and Testing Loss of Video Data in CREMAD Dataset, (d) Training and Testing Accuracy of Video Data in CREMAD Dataset, (e) Training and Testing Loss of Multi Modal Data in CREMAD Dataset, (f) Training and Testing Accuracy of Multi Modal Data in CREMAD Dataset.

TABLE II. ACCURACY AND LOSS OF RAVDESS AND CREMAD DATASETS ON DIFFERENT TYPES OF DATA IN THE DATASET

| Name of Dataset | Type of Data | Train Accuracy | Test Accuracy | Train Loss | Test Loss |
|-----------------|--------------|----------------|---------------|------------|-----------|
| RAVDESS         | Audio        | 92.19          | 95.56         | 0.2430     | 0.161     |
|                 | Video        | 96.92          | 96.28         | 0.1716     | 0.173     |
|                 | Multi Modal  | 91.95          | 95.07         | 0.2155     | 0.136     |
| CREMAD          | Audio        | 72.70          | 80.50         | 0.7293     | 0.588     |
|                 | Video        | 95.33          | 97.88         | 0.1978     | 0.151     |
|                 | Multi Modal  | 65.20          | 69.66         | 0.7840     | 0.646     |

Fig. 5(a) and 5(b) represent the confusion matrix/classification report of how the classes are classified during the testing phase on audio data of the RVDSR and CREMAD datasets. It has been observed that an average precision, recall, f1-score and support values of 0.955, 0.954, 0.953 and 540 on RAVDESS, 0.807, 0.806, 0.805 and 744 on CREMAD datasets, respectively. The detailed description of various emotions and their respective performance measure values of RAVDESS and CREMAD audio data is given in Table III.



(a)

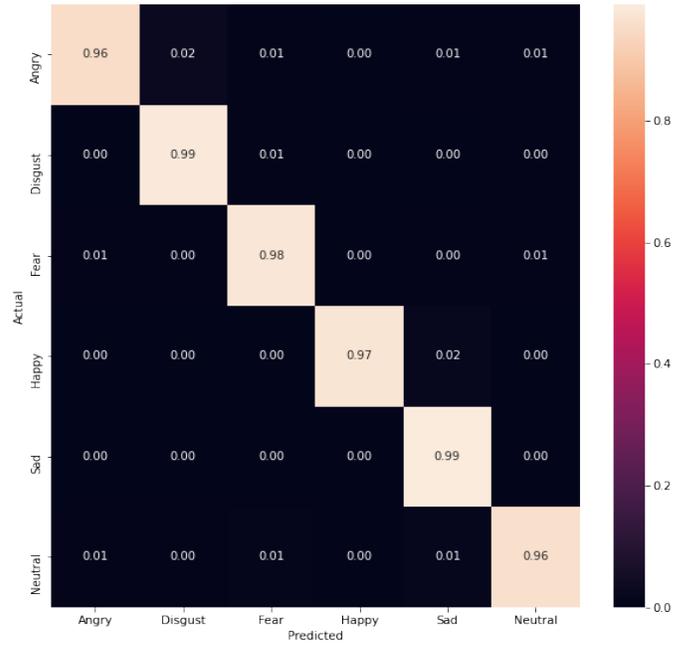
(b)

Fig. 5. (a) Confusion Matrix/Classification Report of RAVDESS Dataset Audio Data, (b) Confusion Matrix/Classification Report of CREMAD Dataset Audio Data.

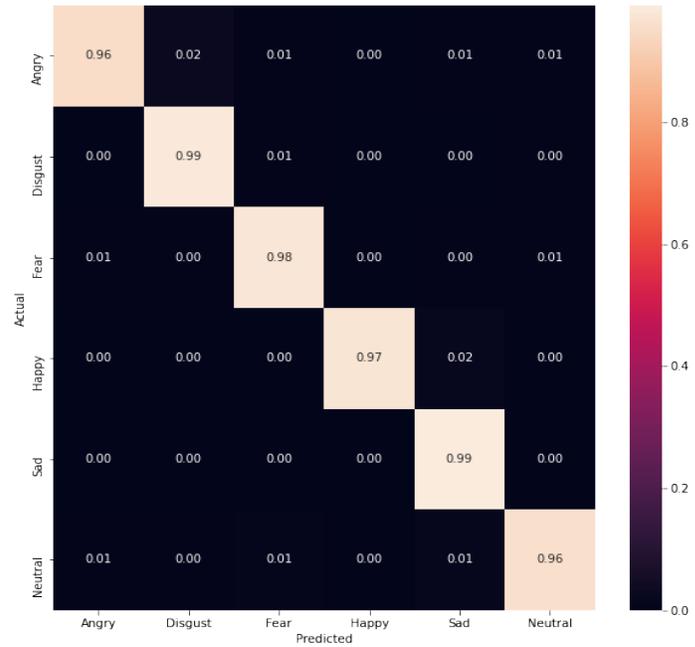
TABLE III. PERFORMANCE METRICS OF RAVDESS AND CREMAD DATASETS ON AUDIO DATA

| Name Of the Dataset | Type of Emotion | Performance Metrics |        |          |         |
|---------------------|-----------------|---------------------|--------|----------|---------|
|                     |                 | Precision           | recall | f1-score | Support |
| RAVDESS             | Angry           | 0.98                | 0.97   | 0.98     | 564     |
|                     | Calm            | 0.97                | 0.97   | 0.97     | 516     |
|                     | Disgust         | 0.98                | 0.92   | 0.95     | 580     |
|                     | Fear            | 0.95                | 0.98   | 0.96     | 644     |
|                     | Happy           | 0.94                | 0.98   | 0.96     | 592     |
|                     | Neutral         | 0.90                | 0.95   | 0.92     | 332     |
|                     | Sad             | 0.97                | 0.90   | 0.93     | 568     |
|                     | Surprise        | 0.95                | 0.96   | 0.95     | 524     |
| CREMAD              | Angry           | 0.94                | 0.87   | 0.90     | 774     |
|                     | Disgust         | 0.75                | 0.82   | 0.78     | 726     |
|                     | Fear            | 0.82                | 0.73   | 0.78     | 748     |
|                     | Happy           | 0.81                | 0.81   | 0.81     | 812     |
|                     | Neutral         | 0.73                | 0.85   | 0.78     | 616     |
|                     | Sad             | 0.79                | 0.76   | 0.78     | 790     |

Fig. 6(a) and 6(b) represent the confusion matrix/classification report of how the classes are classified during the testing phase on video data of the RVDSR and CREMAD datasets. It has been observed that an average precision, recall, f1-score and support values of 0.98, 0.985, 0.985 and 997 on RAVDESS, 0.973, 0.975, 0.975 and 1027 on CREMAD datasets, respectively. The detailed description of various emotions and their respective performance measure values of RAVDESS and CREMAD video data is given in Table IV.



(a)



(b)

Fig. 6. (a) Confusion Matrix/Classification Report of RAVDESS Dataset Video Data, (b) Confusion Matrix/Classification Report of CREMAD Dataset Video Data.

TABLE IV. PERFORMANCE METRICS OF RAVDESS AND CREMAD DATASETS ON VIDEO DATA

| Name Of the Dataset | Type of Emotion | Performance Metrics |        |          |         |
|---------------------|-----------------|---------------------|--------|----------|---------|
|                     |                 | Precision           | recall | f1-score | Support |
| RAVDESS             | Angry           | 0.99                | 0.98   | 0.98     | 1028    |
|                     | Disgust         | 0.97                | 0.99   | 0.99     | 1012    |
|                     | Fear            | 0.98                | 0.99   | 0.98     | 1075    |
|                     | Happy           | 0.99                | 0.99   | 0.99     | 940     |
|                     | Neutral         | 0.97                | 0.99   | 0.99     | 1082    |
|                     | Sad             | 0.98                | 0.97   | 0.98     | 842     |
|                     |                 |                     |        |          |         |
| CREMAD              | Angry           | 0.99                | 0.96   | 0.97     | 1068    |
|                     | Disgust         | 0.96                | 0.99   | 0.98     | 1031    |
|                     | Fear            | 0.97                | 0.98   | 0.97     | 1084    |
|                     | Happy           | 0.99                | 0.97   | 0.98     | 987     |
|                     | Neutral         | 0.96                | 0.99   | 0.98     | 1108    |
|                     | Sad             | 0.97                | 0.96   | 0.97     | 884     |

A multi-modal dataset has been obtained by combing the features of audio and video by using the feature level fusion techniques described in the feature level fusion section of the proposed method on the RAVDESS and CREMAD datasets. Fig. 7(a) and 7(b) give the classification report/confusion matrix obtained from the proposed CNN architecture during the evaluation stage. The classification report shows how the six classes, namely angry, disgust, fear, happy, neutral, and sad, are properly classified during their test by the proposed CNN architecture. An average precision of 0.953 & 0.716, recall of 0.953 & 0.7, f1-support of 0.951 & 0.688, and support of 1613 & 2817 were observed by the proposed CNN architecture during the evaluation phase on RAVDESS & CREMAD multimodal. The detailed description of the results is explained in Table V.

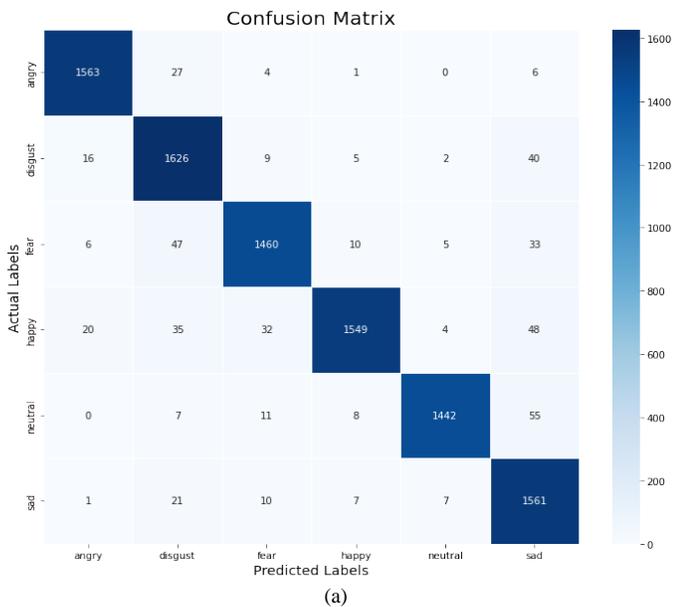


Fig. 7. (a) Confusion Matrix/Classification Report of RAVDESS Dataset on Multimodal Data, (b) Confusion Matrix/Classification Report of CREMAD Dataset on Multimodal Data.

TABLE V. PERFORMANCE METRICS OF RAVDESS AND CREMAD DATASETS ON MULTIMODAL DATA

| Name of the Dataset | Type of Emotion | Performance Metrics |        |          |         |
|---------------------|-----------------|---------------------|--------|----------|---------|
|                     |                 | Precision           | recall | f1-score | Support |
| RAVDESS             | Angry           | 0.97                | 0.98   | 0.97     | 1601    |
|                     | Disgust         | 0.92                | 0.96   | 0.94     | 1698    |
|                     | Fear            | 0.96                | 0.94   | 0.95     | 1561    |
|                     | Happy           | 0.98                | 0.92   | 0.95     | 1688    |
|                     | Neutral         | 0.99                | 0.95   | 0.97     | 1523    |
|                     | Sad             | 0.90                | 0.97   | 0.93     | 1607    |
| CREMAD              | Angry           | 0.56                | 0.25   | 0.35     | 2917    |
|                     | Disgust         | 0.81                | 0.74   | 0.77     | 2836    |
|                     | Fear            | 0.88                | 0.76   | 0.82     | 2891    |
|                     | Happy           | 0.74                | 0.77   | 0.75     | 2621    |
|                     | Neutral         | 0.79                | 0.87   | 0.83     | 2821    |
|                     | Sad             | 0.49                | 0.81   | 0.61     | 2817    |

A detailed description of the macro average and weighted average accuracies Precision, recall, f1-score and support of RAVDESS and CREMAD datasets in all the three modes (Audio, video, and multimoded) are given in Table VI.

The performance of the current work done has been compared with earlier work. It has been observed that the proposed method performed better, and a detailed description of the comparisons is given in Table VII.

TABLE VI. MACRO AVERAGE AND WEIGHTED ACCURACIES OF PERFORMANCE METRICS IN DIFFERENT MODES ON RAVDESS AND CREMAD DATASETS

| Name of The Dataset | Type of Accuracy          | Type of Data | Performance Metrics |        |          |         |
|---------------------|---------------------------|--------------|---------------------|--------|----------|---------|
|                     |                           |              | Precision           | recall | f1-score | Support |
| RAVDESS             | Macro Average Accuracy    | Audio        | 0.95                | 0.96   | 0.95     | 1080    |
|                     |                           | Video        | 0.97                | 0.96   | 0.97     | 9292    |
|                     |                           | Multimodal   | 0.95                | 0.95   | 0.95     | 1578    |
|                     | Weighted Average Accuracy | Audio        | 0.96                | 0.96   | 0.96     | 1080    |
|                     |                           | Video        | 0.97                | 0.97   | 0.97     | 9232    |
|                     |                           | Multimodal   | 0.95                | 0.95   | 0.95     | 1578    |
| CREMAD              | Macro Average Accuracy    | Audio        | 0.81                | 0.81   | 0.80     | 4466    |
|                     |                           | Video        | 0.95                | 0.96   | 0.95     | 12642   |
|                     |                           | Multimodal   | 0.71                | 0.70   | 0.69     | 16903   |
|                     | Weighted Average Accuracy | Audio        | 0.81                | 0.80   | 0.81     | 4466    |
|                     |                           | Video        | 0.97                | 0.96   | 0.96     | 12642   |
|                     |                           | Multimodal   | 0.71                | 0.70   | 0.69     | 16903   |

TABLE VII. ACCURACY COMPARISON OF PROPOSED METHOD WITH ALREADY EXISTING RESULTS

| Name of The Author        | Datasets Used | Percentage of Test Accuracy | Proposed Method Accuracy |
|---------------------------|---------------|-----------------------------|--------------------------|
| Fu, Ziwang, et al. [47]   | RAVDSR        | 75.76                       | 95.07%                   |
| R. Chatterjee et al. [48] |               | 90.48                       |                          |
| Chang X et al. [49]       |               | 91.4                        |                          |
| Wang W et al. [50]        |               | 89.8                        |                          |
| Rory Beard et al [51]     |               | 58.33                       |                          |
| Rory Beard et al [51]     | CREMAD        | 65.0                        | 69.66%                   |
| Ghaleb E et al. [52]      |               | 66.5                        |                          |
| He G et al. [53]          |               | 64                          |                          |

V. CONCLUSION AND FUTURE WORK

A multimodal system for emotion recognition was proposed in the current work. Audio and video information are used here. Audio features are obtained by the Mel-Frequency Cepstral Coefficients extraction technique, and all the videos are converted into images and stored in a spatial-temporal space. The image features are extracted by using a Gaussian weighted function. The MFA fusion technique is to fuse the audio and video features, and the resultant features are given to the FERCNN Model for training and evaluation. For experimentation, the RAVDESS and CREMAD datasets, which consist of audio and video data, are used. Test accuracies of 95.07 and 69.66 were obtained on the RAVDSR

and CREMAD datasets in multimodal mode. Even though many multimodal emotional datasets exist, only two of them are considered. An efficient multimodal system that is generic to all types of multimodal emotional databases can be designed, and the maximum multimodal data accuracy on the CREMAD dataset is 69.66%, which can be further improved.

REFERENCES

- [1] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 975–985, 2019.
- [2] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain, "Multimodal sentiment analysis: Addressing key issues and setting up the baselines," *IEEE Intell. Syst.*, vol. 33, no. 6, pp. 17–25, Nov./Dec. 2018.
- [3] N. Colneric and J. Demsar, "Emotion recognition on Twitter: Comparative study and training a unison model," *IEEE Trans. Affective Comput.*, to be published.
- [4] K. P. Seng and L.-M. Ang, "Video analytics for customer emotion and satisfaction at contact centers," *IEEE Trans. Human-Mach. Syst.*, vol. 48, no. 3, pp. 266–278, May 2017.
- [5] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *Proc. Defense Sci. Res. Conf. Expo (DSR)*, Aug. 2011, pp. 1–5.
- [6] Wang, X., Chen, X., & Cao, C. (2020). Human emotion recognition by optimally fusing facial expression and speech feature. *Signal Processing: Image Communication*, 84, 115831.
- [7] Anbarjafari, G., Noroozi, F., Marjanovic, M., Njegos, A., & Escalera, S. (2019). Audio-Visual Emotion Recognition in Video Clips.
- [8] Srinivas, P. V. V. S., & Mishra, P. (2021). Facial Expression Detection Model of Seven Expression Types Using Hybrid Feature Selection and Deep CNN. In *International Conference on Intelligent and Smart Computing in Data Analytics: ISCD 2020* (pp. 89-101). Springer Singapore.
- [9] Mishra, Pragnyan, and Srinivas, P. V. V. S. "Facial Emotion Recognition Using Deep Convolutional Neural Network and Smoothing, Mixture Filters Applied during Preprocessing Stage." *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 4, 1 Dec. 2021, pp. 889–900., <https://doi.org/10.11591/ijai.v10.i4.pp889-900>.
- [10] P V V S Srinivas and Pragnyan Mishra, "An Improvised Facial Emotion Recognition System using the Optimized Convolutional Neural Network Model with Dropout" *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(7), 2021.
- [11] T. Wu, S. Fu, and G. Yang, "Survey of the facial expression recognition research," in *Proc. Int. Conf. Brain Inspired Cognitive Syst.*, 2012, pp. 392–402.
- [12] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multilevel dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, Mar. 2016.
- [13] Y. N. Chae, T. Han, Y.-H. Seo, and H. S. Yang, "An efficient face detection based on color-filtering and its application to smart devices," *Multimedia Tools Appl.*, vol. 75, pp. 1–20, 2016.
- [14] K. Anderson and P. W. McOwan, "A real-time automated system for the recognition of human facial expressions," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 36, no. 1, pp. 96–105, Feb. 2006.
- [15] Fuad, M., Hasan, T., Fime, A. A., Sikder, D., Iftee, M., Raihan, A., ... & Islam, M. (2021). Recent Advances in Deep Learning Techniques for Face Recognition. *arXiv preprint arXiv:2103.10492*.
- [16] Kamarol, S. K. A., Jaward, M. H., Parkkinen, J., & Parthiban, R. (2016). Spatiotemporal feature extraction for facial expression recognition. *IET Image Processing*, 10(7), 534-541.
- [17] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1319–1329, Jul. 2016. [9].
- [18] S. Nematy and A. R. Naghsh-Nilchi, "Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video

- retrieval,” in Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA), Apr. 2017, pp. 222–228.
- [19] K. P. Seng, L.-M. Ang, and C. S. Ooi, “A combined rule-based & machine learning audio-visual emotion recognition approach,” *IEEE Trans. Affective Comput.*, vol. 9, no. 1, pp. 3–13, Jan./Mar. 2018.
- [20] S. E. Kahou, et al., “EmoNets: Multimodal deep learning approaches for emotion recognition in video,” *J. Multimodal User Interfaces*, vol. 10, pp. 1–13, 2015. [26].
- [21] Y. Wang, L. Guan, and A. N. Venetsanopoulos, “Kernel crossmodal factor analysis for information fusion with application to bimodal emotion recognition,” *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.
- [22] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Inf. Fusion*, vol. 37, pp. 98–125, Sep. 2017.
- [23] E. Cambria, “Affective computing and sentiment analysis,” *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.
- [24] S. K. D’Mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” *ACM Comput. Surv.*, vol. 47, no. 3, 2015, Art. no. 43.
- [25] F. A. Salim, F. Haider, O. Conlan, and S. Luz, “An approach for exploring a video via multimodal feature extraction and user interactions,” *J. Multimodal User Interfaces*, vol. 12, no. 4, pp. 285–296, 2018.
- [26] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, “CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset,” in *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 1 Oct.–Dec. 2014, doi: 10.1109/TAFFC.2014.2336244.
- [27] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE* 13(5): 0196391. <https://doi.org/10.1371/journal.pone.0196391>.
- [28] Jackson, P., & Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.
- [29] F. Cid, L. J. Manso, and P. Nunez, “A novel multimodal emotion recognition approach for affective human robot interaction,” *Proc. FinE*, pp. 1–9, 2015.
- [30] D. Gharavian, M. Bejani, and M. Sheikhan, “Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks,” *Multimedia Tools Appl.*, vol. 76, pp. 1–22, 2016.
- [31] K.-C. Huang, H.-Y. S. Lin, J.-C. Chan, and Y.-H. Kuo, “Learning collaborative decision-making parameters for multimodal emotion recognition,” in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.
- [32] S. Lee, D. K. Han, and H. Ko, “Fusion-convbert: Parallel convolution and bert fusion for speech emotion recognition,” *Sensors*, vol. 20, no. 22, p. 6688, 2020.
- [33] Y. Xu, H. Xu, and J. Zou, “Hgfmm: A hierarchical grained and feature model for acoustic emotion recognition,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6499–6503.
- [34] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, “Audio-visual emotion recognition in video clips,” *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 60–75, Jan./Mar. 2019.
- [35] N. Xu, W. Mao, and G. Chen, “Multi-interactive memory network for aspect based multimodal sentiment analysis,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 371–378.
- [36] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” *arXiv preprint arXiv:1707.07250*, 2017.
- [37] P. P. Liang, Z. Liu, A. Zadeh, and L.-P. Morency, “Multimodal language analysis with recurrent multistage fusion,” *arXiv preprint arXiv:1808.03920*, 2018.
- [38] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, “Efficient low-rank multimodal fusion with modality-specific factors,” *arXiv preprint arXiv:1806.00064*, 2018.
- [39] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.
- [40] D. Ghosal, M. S. Akhtar, D. Chauhan, S. Poria, A. Ekbal, and P. Bhattacharyya, “Contextual inter-modal attention for multi-modal sentiment analysis,” in *proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 3454–3466.
- [41] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, “Learning factorized multimodal representations,” *arXiv preprint arXiv:1806.06176*, 2018.
- [42] Harris, C., Stephens, M.: ‘A combined corner and edge detector’. *Proc. Alvey Vision Conf.*, Manchester, 1988, pp. 147–152.
- [43] Förstner, W., Gülch, E.: ‘A fast operator for detection and precise location of distinct points, corners and centres of circular features. *Proc. ISPRS Inter Commission Conf. Fast Processing of Photogrammetric Data*, Interlaken, June 1987, pp. 281–305.
- [44] S. Zhang, Q. Huang, S. Jiang, W. Gao, and Q. Tian, “Affective visualization and retrieval for music video,” *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 510–522, Oct. 2010.
- [45] N. Fragopanagos and J. G. Taylor, “Emotion recognition in human-computer interaction,” *Neural Netw.*, vol. 18, no. 4, pp. 389–405, 2015.
- [46] Ieymani, M. Pantic, and T. Pun, “Multimodal emotion recognition in response to videos,” *IEEE Trans. Affect. Comput.*, vol. 3, no. 2, pp. 211–223, Apr. 2012.
- [47] Fu, Z., Liu, F., Wang, H., Qi, J., Fu, X., Zhou, A., & Li, Z. (2021). A cross-modal fusion network based on self-attention and residual structure for multimodal emotion recognition. *arXiv preprint arXiv:2111.02172*.
- [48] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra, and D. Giri, “Real-time speech emotion analysis for smart home assistants,” *IEEE Transactions on Consumer Electronics*, vol. 67, no. 1, pp. 68–76, 2021.
- [49] Chang, X., & Skarbek, W. (2021). Multi-Modal Residual Perceptron Network for Audio–Video Emotion Recognition. *Sensors*, 21(16), 5452.
- [50] Wang, W.; Tran, D.; Feiszli, M. What Makes Training Multi-Modal Classification Networks Hard? In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 16–18 June 2020; pp. 12692–12702.
- [51] Beard, R., Das, R., Ng, R. W., Gopalakrishnan, P. K., Eerens, L., Swietojanski, P., & Miksik, O. (2018, October). Multi-modal sequence fusion via recursive attention for emotion recognition. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 251-259).
- [52] Ghaleb, E., Popa, M., & Asteriadis, S. (2019). Metric learning-based multimodal audio-visual emotion recognition. *Ieee Multimedia*, 27(1), 37-48.
- [53] He, G., Liu, X., Fan, F., & You, J. (2020). Image2audio: Facilitating semi-supervised audio emotion recognition with facial expression image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (pp. 912-913).

# BERT based Named Entity Recognition for Automated Hadith Narrator Identification

Emha Taufiq Luthfi<sup>1</sup>  
Faculty of Computer  
Universitas Amikom Yogyakarta  
Yogyakarta, Indonesia

Zeratul Izzah Mohd Yusoh<sup>2</sup>, Burhanuddin Mohd  
Aboobaidar<sup>3</sup>  
Faculty of Information, Communication and Technology  
Universiti Teknikal Malaysia Melaka  
Melaka, Malaysia

**Abstract**—Hadith serves as a second source of Islamic law for Muslims worldwide, especially in Indonesia, which has the world's most significant Muslim population of 228.68 million people. However, not all Hadith texts have been certified and approved for use, and several falsified Hadiths make it challenging to distinguish between authentic and fabricated Hadiths. In terms of Hadith science, determining the authenticity of a Hadith can be accomplished by examining its Sanad and Matn. Sanad is an essential aspect of the Hadith because it indicates the chain of the Narrator who transmits the Hadith. The research reported in this paper provides an advanced Natural Language Processing (NLP) technique for identifying and authenticating the Narrator of Hadith as a part of Sanad, utilizing Named Entity Recognition (NER) to address the necessity of authenticating the Hadith. The NER technique described in the research adds an extra feed-forward classifier to the last layer of the pre-trained BERT model. In the testing process using Cahya/bert-base-indonesian-1.5G, the proposed solution received an overall F1-score of 99.63 percent. On the Hadith Narrator Identification using other Hadith passages, the final examination yielded a 98.27 percent F1-score.

**Keywords**—Hadith narrator; hadith authentication; natural language processing; named entity recognition; NLP; NER; BERT; BERT fine-tune

## I. INTRODUCTION

Islam is a massive faith globally, with over 2 billion Muslims in 2018, accounting for approximately 29.04% of global society. Indonesia has the most massive Muslim community, with more than 233.38 million Muslims in 2018. Muslims need to refer to Islamic rules in the Holy Qur'an and Hadith as their life guidance. The Holy Al-Qur'an is the absolute revelation from Allah (God of Muslims), while Hadith, notably, is a compilation of quotes that Prophet Mohammed has said. Nevertheless, not all the Hadith text is authenticated and authorized to apply. Several fabricated Hadiths cause many issues in determining between genuine and non-genuine Hadiths. The existence of fabricated Hadiths denigrates and reduces the Hadiths' authority and significantly affects Muslim's entire lives, mainly in belief, law, morals, observance, and others. The worst effect of fabricated Hadiths is the confusion they bring to Muslims and, consequently, corrupt their faith [1]. Therefore, it is vital to investigate to verify the authenticity and originality of the accessed Hadiths.

Hadith verification can employ two principal parameters that possibly recognize the condition of a particular Hadith: (1) the context (the meaning of the Hadith itself) and (2) the narrators (the people who recite the Hadith). Recognizing the narrators' names has a crucial role in authorizing a particular Hadith. For example, the snippet of text from Indonesia's Hadith is below:

Telah menceritakan kepada kami Al Humaidi Abdullah bin Az Zubair dia berkata, Telah menceritakan kepada kami Sufyan yang berkata, bahwa Telah menceritakan kepada kami Yahya bin Sa'id Al Anshari berkata, telah mengabarkan kepada kami Muhammad bin Ibrahim At Taimi, bahwa dia pernah mendengar Alqamah bin Waqash Al Laitsi berkata; saya pernah mendengar Umar bin Al Khaththab diatas mimbar berkata; saya mendengar Rasulullah shallallahu'alaihi wasallam bersabda: "Semua perbuatan tergantung niatnya, dan (balasan) bagi tiap-tiap orang (tergantung) apa yang diniatkan; Barangsiapa niat hijrahnya karena dunia yang ingin digapainya atau karena seorang perempuan yang ingin dinikahinya, maka hijrahnya adalah kepada apa dia diniatkan"

The meaning is:

Has told us Al Humaidi Abdullah bin Az Zubair he said, Has told us Sufyan who said, That has said us Yahya bin Sa'id Al Ansari said, has told us Muhammad bin Ibrahim At Taimi, that he had heard of Alqamah bin Waqash Al Laitsi said; I once heard Umar bin Al Khaththab on the pulpit say; I heard the Prophet sallallaahu'alaihi wasallam say: "All actions depend on the intention, and (retribution) for each person (depending on) what is intended; Whoever intends to emigrate because of the world he wants to achieve or because of a woman he wants to marry, then his hijrah is what is he intended for?"

The example Hadith text above has five narrators (highlighted in gray). The status of the above Hadith, authentic or not, can be assessed by identifying and assessing the worthiness of the five narrators. NER is an NLP role that recognizes and classifies named entities in a provided text. "Named entities" refer to predefined semantic categories such as people, locations, and organizations. NER is theoretically applicable to various domains and languages. Therefore, it is challenging to address NER to identify the Hadith Narrator and authenticate it.

This study proposes semi-supervised BERT (Bidirectional Encoder Representations from Transformers) with an extra feed-forward neural network for Hadith Narrators to execute NER, particularly for Indonesian Hadith texts. In case all of the Hadith Narrators have already been identified using the proposed NER Model. Then it is possible to continue with defining the Hadith authentication. The remainder of the essay is organized as follows: To begin, Section II reviews prior work on NER and Hadith Narrator Identification. Section III substantiates this view by discussing the academic definitions of the NER and BERT and the evaluation factors used. Then, Section IV clarifies the recommended model for this investigation. Section V discusses the findings of this research examination. In the end, Section VI discusses the final findings and future research directions.

## II. RELATED WORK

The implementation of NER is domain and language-dependent. When utilized in other domains, the NER generated for one domain performs poorly [2][3]. The scope of the study reported in this paper is limited to identifying the Hadith Narrator using the NER technique. Specifically, Hadith in the Indonesian language.

The study [4] proposes a new Part of Speech (POS) tag and rule-based narrator name extraction for Malay Hadith text. The result was the creation of the POS tag involving 256 words developed from Hadith text, and the rules were created based on five Narrator chains. Similarly, in [2], the author presents a unique rule-based technique for automatically identifying person-name entities in the Malay Hadith text-domain. The model was developed by manually recognizing the names and mannerisms of 150 Malay Hadith books and then developing rules based on them.

The study [5] created a model of NER for Hadith texts written in English. The proposed model makes use of the Support Vector Machine (SVM), the Maximum Entropy Classifier (ME), as well as the Naive Bayes (NB), and classifier combination methods. The results indicate that the classifiers' combination technique achieves the best performance, with precision, recall, and F-Measure values of 96.9 percent, 93.6 percent, and 95.3 percent, respectively. Another author [6] built a NER-based knowledge extraction framework that employs finite-state transducers (FSTs) – KEFST – to extract the Hadith Narrators from the Urdu Translation Hadith text. KEFST consists of five steps: content extraction, tokenization, part of speech tagging, multi-word detection, and NER. This study achieved a precision, recall, and F-measure sequentially of 68%, 75%, and 72%.

The study [7] constructed NERs for Arabic Hadith texts using three machine learning algorithms: naive Bayes, K-nearest Neighbor, and Decision Tree. During the training phase, the NER model achieved a precision of 90% and a recall of 82%. Evaluating the created model on various corpora demonstrates that it can achieve an accuracy of 80% and a recall of 73%. The author [8] constructed NERs for Arabic Hadith texts using three distinct approaches: rule-based, statistical, and hybrid (rule-based combined with statistical). The statistical methods used are the Log-likelihood Ratio (LLR), Point-wise Mutual Information (PMI), S-cost, R-cost,

and U-cost. LLR outperformed PMI, S-cost, R-cost, and U-cost, capturing 76% of the F-measure. Additionally, the rule-based approach captured 80% of the F-measure. The experimental results indicate that the proposed hybrid method of rule-based and statistical analysis achieved an F-measure of 82 percent, which is a positive outcome compared to the individual approach.

The study [9] used two RNN-based models to recognize and categorize named things in Classical Arabic Hadith text by fine-tuning the pre-trained BERT language model. Additionally, this study investigates alternative designs for the BERT-BGRU/BLSTM-CRF models. The BERT-BGRU-CRF model outperformed the other models with an F-measure of 94.76 percent on the CANERCorpus. Another author [10] developed a novel NER model for Arabic Hadith text extracted from the Sahih Bukhari Urdu translation book. The proposed model extracts entities from Hadith text using Finite State Transducers (FST) and subsequently tags them using Conditional Random Fields (CRF). The model had a precision of 96.44 percent, a recall of 88.77 percent, and an F-Measure of 92.41 percent. Similarly, in [11], the author proposes a new approach for extracting Arabic person names from Arabic Hadith text. This study built NER using N-gram phrase extraction and a simple rules model, and the result showed excellent precision of around 84%.

The work [12] offers a novel NER model for Hadith texts in Indonesia using Support Vector Machines (SVM). The results suggest that the NER model attained the most incredible F-1 score of 0.9 using 140 Hadiths containing 1564 entities for training and 60 Hadiths containing 677 entities for testing. Another study [15] built NER with a Naive Bayes classifier for Indonesian Hadith from nine narrators. The results of experiments involving 258 people's names extracted from 13870 tokens of data from 100 Indonesian hadith texts show that combining all features can achieve 82.63% of the F1-Score. The author [13] built NER for indexing names in the Indonesian Hadith Text. This study employs the Hidden Markov Model (HMM). The values of performance that were obtained using HMM's method are 86%. However, by using cross-validation based on the parameters, the performance values increased by 2%, which means that the performance in this research is quite suitable for 38.102 data hadith.

Although various studies have successfully proved the application of NER on Hadith, there is a dearth of studies that optimize BERT to execute NER in order to detect the Narrator in Indonesian Hadith text automatically.

## III. FUNDAMENTAL THEORY

### A. NER

The NER is an NLP task that identifies text fragments related to a specific named entity and categorizes them according to predefined categories such as a person, location, or organization [14]. Four key lines of progress can be seen in the evolution of NER techniques [15]:

- 1) Rule-based techniques are non-annotated and rely on manually written rules.
- 2) Unsupervised learning methods rely on unsupervised

algorithms in the absence of manually labeled training examples.

3) Feature-based techniques for supervised learning depend on supervised learning techniques that have been carefully engineered with features.

4) Deep-learning techniques automatically locate representations required for classification or detection.

Fig. 1 depicts the progress of the NER approach.

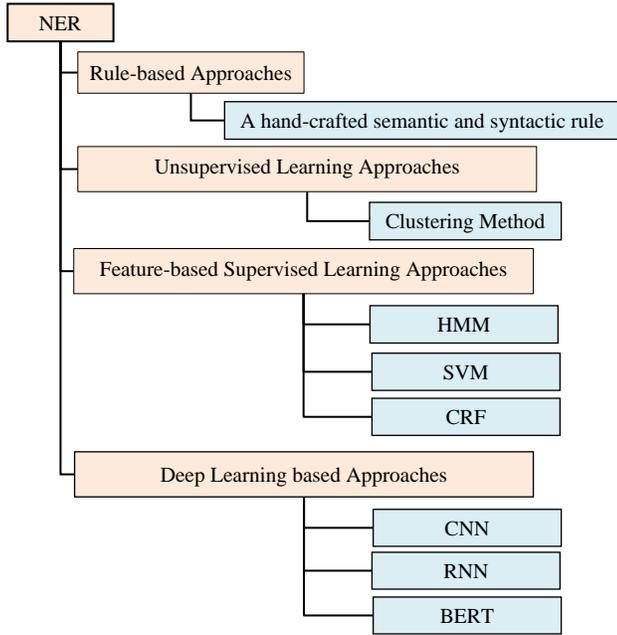


Fig. 1. NER Method's Evolution.

**B. BERT**

A BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer representation that may be fine-tuned to utilize one extra layer of output. BERT enhanced its capacity to create new state-of-the-art outcomes for several assignments, including question answering and sentence categorization, without significantly altering the task-specific architecture [16][17]. Fig. 2 illustrates the BERT architecture.

A feature-based or fine-tuning method might be applied to assign downstream assignments to pre-trained language representations [16]. Fine-tuning is simple, as the transformer's self-attention mechanism enables BERT to perform many downstream assignments on a single text or text pair by swapping the relevant inputs and outputs. Each assignment needs BERT to receive just the assignment-specific inputs and outputs, which fine-tunes all parameters end-to-end.

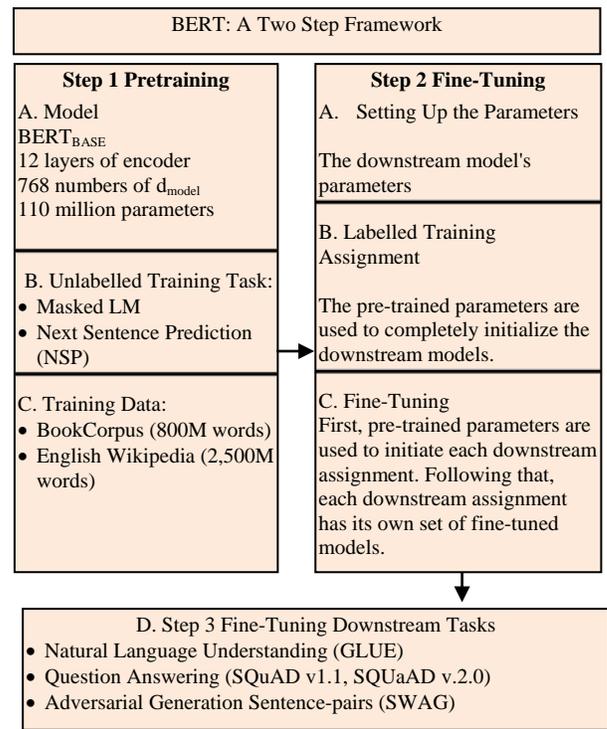


Fig. 2. NER Method's Evolution.

**IV. PRESENTED MODEL**

The methodology used in this investigation is summarized in Fig. 3.

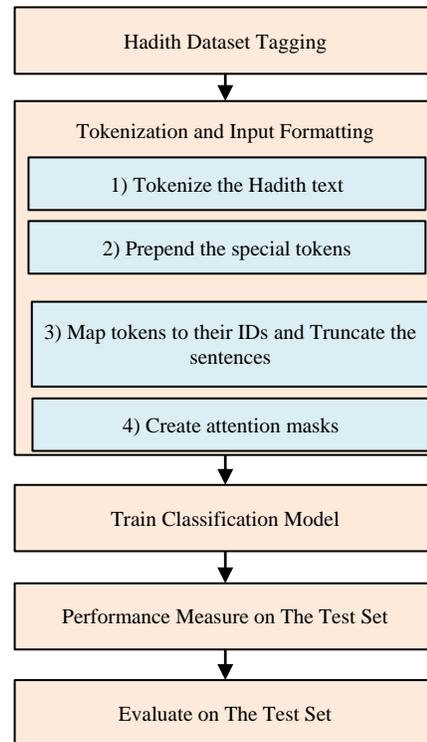


Fig. 3. Proposed Methodology.



TABLE II. SAMPLE OF TOKENS

| WORD        | TOKENS             | WORD         | TOKENS       |
|-------------|--------------------|--------------|--------------|
| dari        | dari               | kami         | kami         |
| abu         | abu                | sesungguhnya | sesungguhnya |
| abdurrahman | abdurrahman        | penciptaan   | penciptaan   |
| bin         | bin                | kalian       | kalian       |
| mas'ud      | mas<br>[UNK]<br>ud | dikumpulkan  | dikumpulkan  |
| ia          | ia                 | dalam        | dalam        |
| berkata     | berkata            | rahim        | rahim        |
| bahwa       | bahwa              | sang         | sang         |
| rasulullah  | rasulullah         | ibu          | ibu          |
| yang        | yang               | selama       | selama       |
| jujur       | jujur              | empat        | empat        |
| dan         | dan                | puluh        | puluh        |
| terpercaya  | terper<br>##caya   | hari         | hari         |
| bersabda    | bersabda           | berupa       | berupa       |
| kepada      | kepada             | sperma       | sperma       |

The number of words that must be tokenized down to the subword level and the final number of tokens for a Hadith text is determined by the used pre-defined BERT. Since the BERT vocabulary for each of the pre-defined BERTs is diverse.

Table III compares the tokenized outcomes of numerous pre-defined BERTs. Most pre-defined BERTs built on the Indonesian language have shorter median token lengths than pre-defined BERTs "bert-base-uncased" built on the general language.

2) *Prepend the special tokens*: Tokenization in BERT entails putting the unique [CLS] token at the beginning of the Hadith text and appending the [SEP] token at the end to indicate the text's beginning and end. This treatment is depicted in Fig. 4 Step C.

3) *Map tokens to their IDs and truncate the sentences*: The word tokens must be mapped to their BERT vocabulary IDs, and all sentences must be made to have the same number of tokens. In order for the GPU to operate on a batch. This process is addressed with some steps, i.e., (1) defining the max sentence length, (2) adding the special [PAD] token to the sentences with the token shorter than the max length, and (3) truncating sentences that are longer than the max length.

The max length adjustment in this study refers to the column max length in Table III, except when employing "bert-base-uncased," in which case the max length is set to 512. Since the limit is derived from the Transformer architecture's positional embeddings, a maximum length must be imposed. Fig. 2 Step D depicts this treatment.

TABLE III. TOKENIZATION RESULTS OF SEVERAL PRE-DEFINED BERT

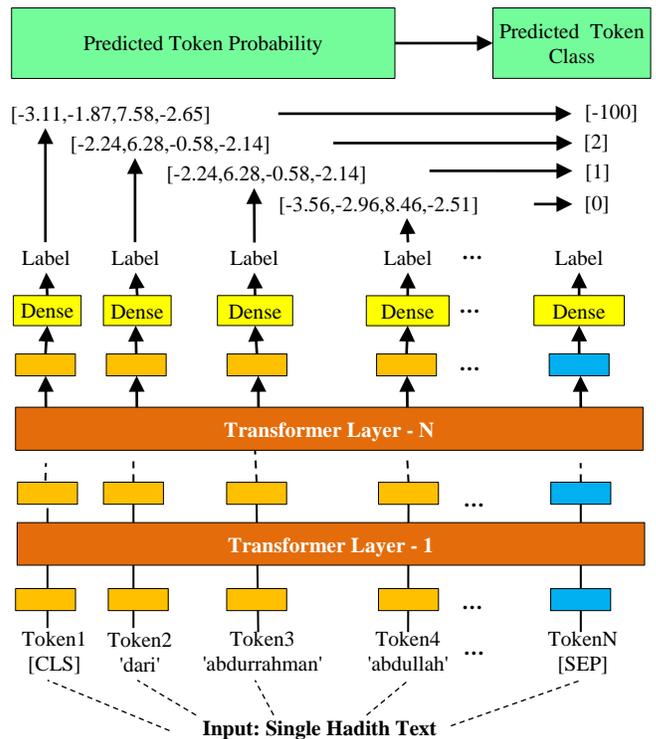
| BERT PRE-TRAINED                | MIN LENGTH | MAX LENGTH | MEDIAN LENGTH |
|---------------------------------|------------|------------|---------------|
| bert-base-uncased               | 90         | 736        | 214           |
| cahya/bert-base-indonesian-1.5G | 50         | 408        | 113           |
| indobenchmark/indobert-base-p1  | 46         | 361        | 104           |
| indobenchmark/indobert-base-p2  | 46         | 361        | 104           |

The max length adjustment in this study refers to the column max length in Table III, except when employing "bert-base-uncased," in which case the max length is set to 512. Since the limit is derived from the Transformer architecture's positional embeddings, a maximum length must be imposed. Fig. 2 Step D depicts this treatment.

4) *Create attention masks*: The final tokenization and formatting process provides the model with an "attention mask" for each sample that identifies and instructs BERT to ignore the [PAD] tokens. This procedure is depicted in Fig. 4 Step E.

C. Classification and Model Training

Thirdly, the classification model must be trained. The proposed model architecture is the BERT with a single linear layer for classifying the entity classes associated with each Narrator token. The proposed model architecture is depicted in Fig. 5.



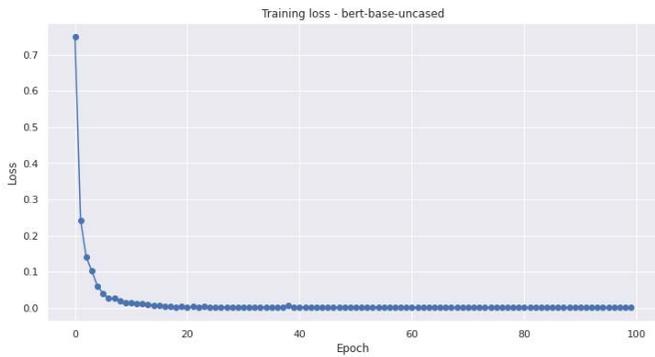
*Dari Abu Abdurrahman Abdullah bin Mas'ud, ia berkata bahwa Rasulullah yang jujur dan terpercaya bersabda kepada kami, "Sesungguhnya, penciptaan kalian dikumpulkan dalam rahim sang ibu selama empat puluh hari berupa sperma..."*

Fig. 5. Classification Model for Hadith Text Tokens.

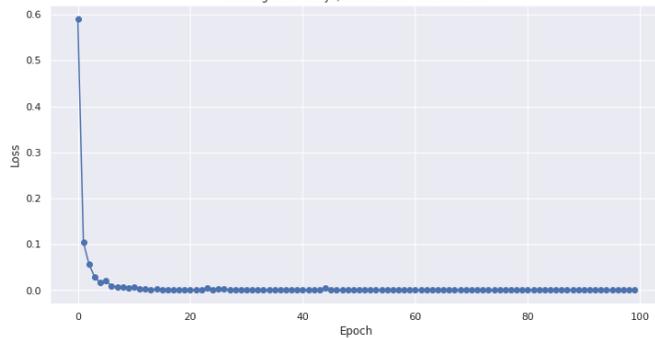
According to the pre-trained model used, the BERT model employs N-Layer Transformers. The final additional layer will consist of two steps for classifying the token into the token class. Forecast probability of tokens first, followed by the class of tokens. This study examines how four different BERT pre-trained models can be scaled up to become NER models. The NER model is the target model, which was trained using the following parameters:

- Maximum Length : 512
- Number of Batches : 16
- Number of Epochs : 100
- Classification Labels 2 for 'B-Narrator', 0 for 'I-Narrator', and 1 for 'O'.

Fig. 6 depicts the training loss graphics associated with the process of fine-tuning the NER Model.



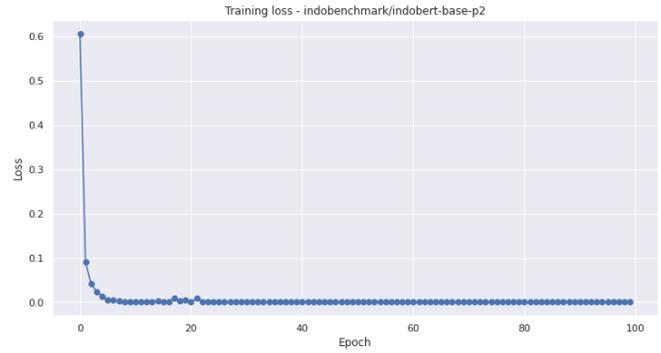
(a) Bert-Base-Uncased



(b) Cahya/Bert-Base-Indonesian-1.5G



(c) Indobenchmark/Indobert-Base-p1



(d) Indobenchmark/Indobert-Base-p2

Fig. 6. The NER Model's Training Loss During the Training Process.

#### D. Performance Measure

The fourth stage is to evaluate the performance of the test set. The performance of the NER Model is quantified using an F1-Score. The F1-score is a numerical term that denotes the compatible mean of precision and recall. The formula below expresses the precision, recall, and F1-score:

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3)$$

The F1-score result for the NER model constructed on top of four separate BERT pre-trained models is shown in Table IV. Except for the bert-base-uncased model, the general BERT model, the BERT pre-trained model contains options supporting the Indonesian language.

TABLE IV. THE NER MODEL F1-SCORE ON TRAINING PROCESS

| BERT PRE-TRAINED                | F1 score |
|---------------------------------|----------|
| bert-base-uncased               | 99.08%   |
| cahya/bert-base-indonesian-1.5G | 99.63%   |
| indobenchmark/indobert-base-p1  | 99.43%   |
| indobenchmark/indobert-base-p2  | 99.11%   |

#### E. Evaluation

The fifth stage involves the evaluation of the NER model. As shown in Table IV, the NER model's performance was assessed again using an additional forty New Hadith texts. The evaluation conclusions are summarized in Table V. The NER model's assessment results are, on average, 0.28 percent lower than the training results. The sequence, on the other hand, is preserved. Cahya/bert-base-indonesian-1.5G demonstrated the best performance, followed by indobenchmark/indobert-base-p1, indobenchmark/indobert-base-p2, and bert-base-uncased.

Table VI summarizes the results of the NER evaluations of three Hadith passages. On hadith1, hadith2, and hadith 3, texts of assessment, some errors in tagging output were made. For instance, in Hadith 1, "Abu" should be labeled as I-Narrator rather than B-Narrator. On hadith2, the word "Ummu" should be categorized as I-Narrator but tagged as "O", and so forth.

TABLE V. THE NER MODEL F1-SCORE ON EVALUATION PROCESS

| BERT PRE-TRAINED                | F1 score |
|---------------------------------|----------|
| bert-base-uncased               | 98.68%   |
| cahya/bert-base-indonesian-1.5G | 99.27%   |
| indobenchmark/indobert-base-p1  | 99.23%   |
| indobenchmark/indobert-base-p2  | 98.94%   |

TABLE VI. THE NER MODEL EVALUATION SAMPLES

| Input   |           | Expected Output | Bert-base-uncased Output | Cahya/bert-base-indonesian-1.5G |
|---------|-----------|-----------------|--------------------------|---------------------------------|
| Tokens  |           |                 |                          |                                 |
| Hadith1 | Dari      | O               | O                        | O                               |
|         | Amirul    | B-Narrator      | B-Narrator               | B-Narrator                      |
|         | Mukminin  | I-Narrator      | I-Narrator               | I-Narrator                      |
|         | Abu       | I-Narrator      | B-Narrator (False)       | I-Narrator                      |
|         | Hafsh     | I-Narrator      | I-Narrator               | I-Narrator                      |
|         | Umar      | I-Narrator      | I-Narrator               | I-Narrator                      |
|         | bin       | I-Narrator      | I-Narrator               | I-Narrator                      |
|         | Khaththab | I-Narrator      | I-Narrator               | I-Narrator                      |
|         | ...       | ...             | ...                      | ...                             |
| Hadith2 | Dari      | O               | O                        | O                               |
|         | Ummul     | B-Narrator      | B-Narrator               | B-Narrator                      |
|         | Mukminin  | I-Narrator      | I-Narrator               | I-Narrator                      |
|         | Ummu      | I-Narrator      | O (False)                | I-Narrator                      |
|         | Abdillah  | I-Narrator      | B-Narrator (False)       | I-Narrator                      |
|         | Aisyah    | I-Narrator      | I-Narrator               | O                               |
|         | ...       | ...             | ...                      | ...                             |
| Hadith3 | syubhat   | O               | B-Narrator (False)       | O                               |
|         | yang      | O               | O                        | O                               |
|         | ...       | ...             | ...                      | ...                             |
|         | menjaga   | O               | O                        | O                               |
|         | dirinya   | O               | O                        | O                               |
|         | dari      | O               | O                        | O                               |
|         | halhal    | O               | B-Narrator (False)       | O                               |
|         | syubhat   | O               | B-Narrator (False)       | O                               |

## V. RESULT

As mentioned previously, NER's implementation is domain and language dependent. When a NER developed for one domain is used in another, it performs poorly [4][5]. This study aims to identify the Hadith Narrator using the NER approach. More precisely, Hadith in Indonesian. The results of this study can be compared to those of the previous study, as indicated in Table VII. Most studies on NER measure its performance by utilizing the F1-score. The distribution of each NER Tag cannot be predicted and may have imbalanced data. An F1-score is needed to capture the harmonic mean of precision and recall. The highest F1-score of the proposed NER model indicates a high value for both precision and recall. The proposed NER model achieved 99.27% of the F1-score.

TABLE VII. COMPARISON OF THE NER MODEL'S PERFORMANCE

| Study      | Dataset           | Methods            | F1-Score |
|------------|-------------------|--------------------|----------|
| [5]        | English Hadith    | SVM                | 95.30%   |
| [6]        | Urdu Hadith       | FST                | 72.00%   |
| [7]        | Arabic Hadith     | NB, KNN, DT        | 86.00%   |
| [8]        | Arabic Hadith     | Rule-based and LLR | 82%      |
| [9]        | Arabic Hadith     | BERT               | 94.76%   |
| [10]       | Arabic Hadith     | FST and CRT        | 92.41%   |
| [11]       | Arabic Hadith     | N-Gram             | 84.00%   |
| [12]       | Indonesian Hadith | SVM                | 90.00%   |
| [18]       | Indonesian Hadith | NB                 | 82.63%   |
| [13]       | Indonesian Hadith | HMM                | 86.00%   |
| This study | Indonesian Hadith | BERT fine-tuned    | 99.27%   |

## VI. CONCLUSION AND FUTURE WORK

The BERT is designed and implemented in this research to provide a NER Hadith Narrator identification using an extra feed-forward classifier. Cahya/bert-base-indonesian-1.5G received the highest F1-score of 99.63 percent during the training phase. On the Hadith Narrator Identification using other Hadith passages, the final examination yielded a 98.27 percent F1-score. It suggests that when utilized to identify Hadith Narrators for Indonesian Hadith texts, the suggested NER model in this work performs best.

There are several future avenues that this experiment should take. The first step in developing an experiment dataset is to increase the amount of Hadith texts. This is required since the Hadith contains a range of Sanad and Matn forms. The current study then applies three IOB tags, namely B-Narrator, I-Narrator, and O. These three tags sufficed because all that was required was to determine which words were associated with the Narrator and which were not. Additional study is warranted in light of the Hadith dataset's addition. Other tags must be considered, such as those for Rasulullah, Taraf, and others.

REFERENCES

- [1] A. H. Usman and R. Wazir, "The Fabricated Hadith: Islamic Ethics and Guidelines of Hadith Dispersion in Social Media," *Turkish Online J. Des. Art Commun.*, vol. 8, pp. 804–808, 2018.
- [2] N. Abd Rahman, N. Alias, N. K. Ismail, Z. Bin Mohamed Nor, and M. N. B. Alias, "An Identification of Authentic Narrator's Name Features in Malay Hadith Texts," in *ICOS 2015 - 2015 IEEE Conference on Open Systems*, 2015, pp. 79–84.
- [3] C. Y. Lim, I. K. T. Tan, and B. Selvaretnam, "Domain-General Versus Domain-Specific Named Entity Recognition: A Case Study Using TEXT," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11909 LNAI, no. October 2019, pp. 238–246, 2019.
- [4] N. A. Rahman, N. K. Ismail, Z. M. Nor, M. N. Alias, M. S. Kamis, and N. Alias, "Tagging narrator's names in Hadith text," *J. Fundam. Appl. Sci.*, vol. 9, no. 5S, p. 295, 2018.
- [5] M. J. Jaber and S. Saad, "NER In English Translation Of Hadith Documents Using Classifiers Combination," *J. Theor. Appl. Inf. Technol.*, vol. 84, no. 3, pp. 348–354, 2016.
- [6] A. Mahmood, H. U. Khan, Z. Ur Rehman, and M. S. Faisal, "KEFST : a knowledge extraction framework using finite-state transducers," *Electron. Libr.*, vol. 37, no. 2, pp. 365–384, 2019.
- [7] M. A. Siddiqui, M. E. Saleh, and A. A. Bagais, "Extraction and Visualization of the Chain of Narrators from Hadiths using Named Entity Recognition and Classification," *Int. J. Comput. Linguist. Res.*, vol. 5, no. 1, pp. 14–25, 2014.
- [8] S. S. Balgasem and L. Q. Zakaria, "A hybrid method of rule-based approach and statistical measures for recognizing narrators name in hadith," in *Proceedings of the 2017 6th International Conference on Electrical Engineering and Informatics: Sustainable Society Through Digital Innovation, ICEEI 2017*, 2018, vol. March 2018, pp. 1–5.
- [9] N. Alsaaran and M. Alrabiah, "Classical Arabic Named Entity Recognition Using Variant Deep Neural Network Architectures and BERT," *IEEE Access*, vol. 9, pp. 91537–91547, 2021.
- [10] A. Mahmood, H. U. Khan, Zahoor-Ur-Rehman, and W. Khan, "Query based information retrieval and knowledge extraction using Hadith datasets," in *Proceedings - 2017 13th International Conference on Emerging Technologies, ICET2017*, 2018, vol. Feb 2018, pp. 1–6.
- [11] M. Alhawarat, "A domain-based approach to extract Arabic person names using N-grams and simple rules," *Asian J. Inf. Technol.*, vol. 14, no. 8, pp. 287–293, 2015.
- [12] F. A. Yusup, M. A. Bijaksana, and A. F. Huda, "Narrator's name recognition with support vector machine for indexing Indonesian hadith translations," *Procedia Comput. Sci.*, vol. 157, pp. 191–198, 2019.
- [13] W. P. Sari, M. A. Bijaksana, and A. F. Huda, "Indexing Name in Hadith Translation Using Hidden Markov Model ( HMM );" in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, 2019, pp. 1–5.
- [14] A. Zahra, A. F. Hidayatullah, and S. Rani, "Kajian Literatur Named Entity Recognition pada Domain Wisata," *Automata*, vol. 2, no. 1, pp. 0–4, 2021.
- [15] J. Li, A. Sun, J. Han, and C. Li, "A Survey on Deep Learning for Named Entity Recognition," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 1, pp. 1–1, 2020.
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Tautanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.
- [17] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," 2020.
- [18] F. Y. Azalia, M. A. Bijaksana, and A. F. Huda, "Name indexing in Indonesian translation of hadith using named entity recognition with naïve bayes classifier," *Procedia Comput. Sci.*, vol. 157, pp. 142–149, 2019.

# An Early Intervention Technique for At-Risk Prediction of Higher Education Students in Cloud-based Virtual Learning Environment using Classification Algorithms during COVID-19

Dr.Arul Leena Rose.P.J, Ananthi Claral Mary.T\*

Department of Computer Science

College of Science and Humanities, SRM Institute of Science and Technology  
Chengalpattu, India

**Abstract**—Higher Education is considered vital for societal development. It leads to many benefits including a prosperous career and financial security. Virtual learning through cloud platforms has become fashionable as it is expediency and flexible to students. New student learning models and prediction outcomes can be developed by using these platforms. The appliance of machine learning techniques in identifying students at-risk is a challenging and concerning factor in virtual learning environment. When there are few students, it is easy for identification, but it is impractical on larger number of students. This study included 530 higher education students from various regions in India and the outcomes generated from online survey data were analyzed. The main objective of this research is to predict early identification of students at-risk in cloud virtual learning environment by analyzing their demographic characteristics, previous academic achievement, learning behavior, device type, mode of access, connectivity, self-efficacy, cloud platform usage, readiness and effectiveness in participating online sessions using four machine learning algorithms namely K Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Random Forest (RF). Predictive system helps to provide solutions to low performance students. It has been implemented on real data of students from higher education who perform various courses in virtual learning environment. Deep analysis is performed to estimate the at-risk students. The experimental results exhibited that random forest achieved higher accuracy of 88.61% compared to other algorithms.

**Keywords**—Prediction; at-risk; machine learning; virtual learning environment; cloud platforms; classification; COVID-19; random forest; student academic performance

## I. INTRODUCTION

During COVID-19 crisis, the entire education system all over the world has shifted towards virtual learning. Online teaching is highly dependent on the successful delivery of the content. Cloud-based Virtual Learning Environment (VLE) is vital component of education in university environments. This interactive platform enables learners to achieve education objectives during pandemic outbreak. The growth of cloud computing technology has brought new opportunities in the field of education as it facilitates effective and efficient learning mechanism. Benefits of cloud computing includes

collaborative learning environment, expense reduction, scalability, shareable content, usability, and global education. This induces the way online learning can be shared and distributed on diverse types of devices and platforms. Some of the biggest organizations including Amazon, Google, Microsoft, and Oracle are selecting the cloud due to its several benefits. Currently, due to the rising of educational institutions every day there exists a gap between industrial requirements and educational institutions. The technological enhancement of cloud computing might fill the gap by rendering free or paid training to the students' using systems that do not require any additional cost. For instance, educators can widen most crucial sections, upload necessary audio/video materials to support contents in cloud-based platforms, etc. However, despite of these advantages, students' learning behaviors and interaction with digital contents is still limited. Hence, this research analyzes at-risk students in cloud-based virtual learning platforms. Early prediction using classification techniques is an efficient and significant way to deliver timely intervention for dropout.

Many researchers have investigated that there are numerous ways of applying machine learning algorithms in education field. One of the vital focuses is to predict at-risk students in VLE by observing diverse student's attributes. Compared with traditional face-to-face education, dropout rates are higher as there was a sudden transformation towards virtual learning. COVID-19 pandemic has led to increase in digitalization and triggered online learning implementation in education sectors. All over the globe, learning sectors are focusing on virtual learning platforms to enhance the procedure of enlightening students. The effectiveness of online medium depends on students. The predominant usage of VLE has helped them to complete their course work examinations. However, evaluating student's results and predicting at-risk students is complex. Personalized assistance must be given to the students at-risk. In this context, this research focuses on identifying exact predictive models to investigate unidentified information regarding students' demographics characteristics, previous academic history, learning behavior, device type, mode of access and connectivity for online classes, self-efficacy, cloud platform usage, readiness, and effectiveness of participation in virtual

\*Corresponding Author

classes. We have examined that these are the appropriate features that makes an impact of student learning and helps in prediction.

Recently, for schools and higher education sectors online learning has emerged as a vital source and will prolong in future. In virtual learning courses, the collaboration between learners and educators is mediated by virtual learning environment. Chatbox helps the students to interact with instructors and take part in the classroom discussions. Machine learning is widely used in education sectors for classroom management, scheduling, etc. It is a method of personalized learning that provides an individualized educational experience to students where they can manage their own learning, at their own pace, make their own decisions about what to learn. In a classroom using personalized learning, students select what they are interested in, and educators fit the curriculum and guidelines to students' interests. Scheduling helps in searching for an optimal and adaptive teaching policy that helps students learn more efficiently. Dynamic scheduling matches students requiring assistance with teachers allocating time. It is regarding designing algorithms that automatically bring out valuable information from the given data. While machine learning has many success stories, there are software available to design and train rich and flexible machine learning systems. The mathematical foundations are important to build complicated machine learning systems. These models can be created by using diverse features of student data. The machine learning algorithms can be used to develop successful classifiers.

## II. LITERATURE REVIEW

Three determinant factors based on technological, organizational, and environmental contexts impact the acceptance of cloud computing in higher education sectors. Survey method was used to collect data from respondents and powerful statistical tool SmartPLS examined the significance of each of these influencers. Results show that factors of compatibility, security, peak management support, authoritarian policy and relative advantage have positive effect [1]. Dataset was chosen from UK Open University, which enclosed students' activity logs, assignment, and final marks, all stored in VLE logs. Feature selection was essential for designing accurate prediction models thereby facilitating students to improve their performance. A predictive model was constructed by the researchers to early forecast students at-risk of dropout. SELogisticRegression and Input-Output Hidden Markov Model (IOHMM) outperformed other baseline models. The overall accuracy of their proposed model is 84% [2]. Almajali & Masadeh ascertained how enabling circumstances, social media, comfort of utilization affect students' opinion for online learning during COVID-19. Their study proved that enabling conditions had a positive impact. Furthermore, they have discussed about the difficulties faced by the students during pandemic like anxiety, lack of device, issues in internet connectivity, etc. Their findings revealed that students who are expertise in utilization of online learning technologies have positive perception towards it [3]. In educational institutions identifying at-risk students was a problematic task. Naive Bayes classifier was selected for the progress of early warning system. Four classification colors

were used to represent the various warning levels namely green for non-at-risk, yellow represents possible at-risk, red for at-risk and black color indicates dropout based on students' grades [4]. The study employs dataset from two academic writing courses in Hong Kong University. Logistic regression and classification trees can be utilized in higher education perspective, but ANN is not appropriate. The research team suggested that accuracy can be improved with other datasets [5]. The researchers have used extremely limited student attributes. Their predictions help to classify the students who are potentially at-risk. This information facilitates tutors to make timely intervention to increase student success. Three machine learning techniques DT, KNN, RF were used. Static attributes namely sexual category, age and previous educational results were excluded [6].

During pandemic contentment of higher education students aspired for learning in virtual modality was forecasted. Students responded that they obtain their classes, seminars, videoconferences using Google Meet and Zoom compared to WebEx and Blackboard. Researchers have concluded that students have better opinion towards online learning, the difficulty do not stretch out in technology usage, but it lies in the teachers' teaching strategy [7]. A tool that allows estimating the hazard of quitting an academic course was proposed. Python programming language was used to implement several machine learning algorithms namely LDA, SVM and RF. LDA and SVM was proved to have utmost performance with a slight superior variance for SVM results. When additional learning requirements feature were introduced, in random forest, the final performance was improved compared to LDA and SVM results. The suggestion for the future development is to have more data regarding student performance by considering the outcome of their activities completed in virtual education environments namely Moodle, Google Classroom and Edmodo [8]. Francis Ofori et al. reviewed the literature and summarized the various machine learning models with their corresponding prediction accuracy. It helps to improve the graduation rates by providing feedbacks to educators and students thereby modifying learning environments. They concluded that most machine learning models dwelled in studying students' performance prediction but failed to identify the best model [9]. Identifying students' at-risk by using eBook interaction logs have employed 13 prediction algorithms. Results revealed that random forest performed better than other algorithms with accuracy 82.3% and kappa 64.7% for raw data, J48 algorithm with accuracy 83.3% and kappa 66.5% performed better with transformed data. Naive Bayes accuracy 81.1% and kappa 62.1% outperformed other algorithms for categorical data [10]. Automated machine learning usage was proposed to improve correctness of prediction percentage depending on the data before commencing of their academic year. This aids students to moderate their risk failure and has achieved the overall accuracy of 75.9% [11]. Perceptions of post graduate students towards responding online learning process have been discussed. Majority of the students used Zoom cloud platform for learning from home activities. Others preferred Google Classroom, Whatsapp and other applications based on the agreement provided by each platform. Few students declared that limitations in technology and usage of

applications hampered the online learning process [12]. The data of 2097 students of higher education were investigated, and the system was trained with Logistic Regression and ANN with four attributes related to student socio-economic and academic details. Results have shown that highest risk of dropping out of students has lowest grade. Comparing to Logistic Regression, ANN has better classification accuracy. However, the accuracy of ANN did not exceed 80% [13].

The researchers developed an early identification system using student performance and administrative data from private and state university. AdaBoost algorithm was used to predict the student dropout. The results revealed that prediction accuracy improves at fourth semester when compared to first semester. The demographic data available at the time of enrollment does not improve prediction accuracy when performance data is available in increasing semesters; this issue must be solved [14]. Khadija Alhumaid focused on the fear of technology usage by students and educators during Covid-19 pandemic. The various fear factors are uncertainty, anxiety, and fear of losing loved ones. Hence m-learning was adopted by the educational institutions, the results of studying and teaching was promising. She has concluded that with the assistance of mobile learning the fear factors can be reduced. It has a high perceived usefulness and perceived ease of use that can decrease the fear and enable the respondents to achieve their classes on time [15]. Reduced Training Vector-SVM was proposed to predict marginal and at-risk students. Analysis revealed that this algorithm can diminish number of training vectors and training time of classifier by at least 60% thereby retaining accurateness. Results represented an overall accuracy of 93.5% [16]. Two open-source datasets namely mathematics and Portuguese was selected for predicting student educational performance. It was identified that prior grade has most impact on finishing grade. Random Forest has gained higher accuracy in mathematics dataset. SVM attained better accuracy than Random Forest in Portuguese dataset [17]. The four main difficulties of students in virtual learning classes were meager internet connection, lack of experience to virtual education applications and tools namely Teams, Padlet, Socrative and Miro, students' restricted English proficiency and difficulty in concentration. Besides, results have proved that online learning is cost effective, as it's not necessary for students to arrive to campus. They had better time management and have utilized digital devices to access online classes. The students get encouraged when they had quizzes established through interactive applications namely Kahoot, Padlet, Micro, Socrative and many other. Lecturers gave more assignments during online learning period than traditional face-face learning [18]. The troubles faced by at-risk students were analyzed in VLE. These predictive models can be used for avoiding student dropouts. Feature engineering was used to enhance performance of predictive models. Experimental results revealed that random forest provides excellent results when compared to other baseline models [19].

Several machine learning algorithms were employed to the dataset to predict low-engagement students in web-centered learning systems from log data of VLE. Kappa and accuracy values were compared for the models. Outcomes proved that J48, decision tree, JRIP, gradient-boosted classifiers revealed

improved results [20]. Predicting students' difficulties in digital design course session were investigated. SVM has achieved 80% performance accuracy compared to other classifiers namely ANN, LR, NBC, DT for predicting student obscurity [21]. Deep long short-term memory model was deployed for students' performance prediction. This model tends to monitor the week-wise pattern of students' interaction and their engagement activities to learn their behavior and generate better outcomes. It outperformed other baseline models namely logistic regression, ANN. It predicted with 90% accuracy of student interaction in VLE [22]. An application was implemented that utilizes academic information of university students and generated classification models by using ANN, ID3 and C4.5. Decision tree built by C4.5 has higher performance measure. Suggestions were given to increase the number of variables and including the institutional and socio-economic variables for further research [23]. The ways to measure fairness in VLE was examined. CGPA was referred as main attribute for student performance prediction. A great underrepresentation of disabled students was determined. This leads to misclassification that disable students were predicted to fail the course. These guidelines must be considered by the researchers [24]. Zulherman analyzed the strength, weakness, opportunities, and threats to Zoom Cloud Meeting, a reliable video platform. The results focused behavioral intention drivers of ZCM usage during crisis are hedonic inspiration and perceived self-efficacy. Influence among these attributes was stronger [25]. Moreover, researchers focused on predicting student's dropout at course level in e-learning course using various machine learning techniques. Five attributes reflecting course activities namely accesses, assignments, tests, exam, and project were considered. Pearson correlation was conducted to identify correlations between independent variables and results. The most appropriate attribute for prediction with high correlation besides project and tests is access that detects students who do not obtain time for accesses that expose them to higher risk. Most adopted algorithms are Logistic Regression, Decision Tree, Naive Bayes Classifier, Support Vector Machine, Random Forest, Neural Network. Results proved that Random Forest classifier obtained best accuracy with 93%, precision reached 86%, F1 score was 91% compared to other classifiers [26]. Researchers have collected real students' data with various information namely personal, economic, and academic records and evaluated by statistics values to find most effective one. They used three prediction algorithms Decision tree, SVM and KNN to detect student dropout. Decision tree reaches better performance for identifying students at high risk. In decision tree they have used J48 technique that represents real dropout groups. Authors have stated that it is vital to evaluate student behavior through multi objective algorithms that assess their skills and emotions [27]. On the other hand, dropout rates are higher in online learning than offline as students must control their own study time without the help of educators. It is vital for professors to assist students in a timely intervention to avoid dropout. Lowering dropout rate is an important challenge for universities. The experiment collected actual log and historical records from Cyber University Learning Management System (LMS) to predict drop-out risk during learning period. They have used four

machine learning algorithms Decision Tree, Random Forest, SVM and Deep Neural Network. Random Forest shows the best performance with 96% accuracy [28].

In previous studies it is evident that predicting student dropout is challenging task. When different machine learning algorithms are employed, it reveals varying prediction accuracy of students' at-risk. It is inexplicit that which algorithm is most excellent for predicting at-risk students in virtual learning environments, and what are the most appropriate features to be considered for various machine learning classifiers, as the determinants of attrition depend on multi-dimensional character. Also, during time of registering for online classes, the student data collected at the university are not sufficient for dropout prediction. In this scenario, our research focuses on several factors namely prior academic achievement, demographic characteristics, learning behavior, device type, access, connectivity, self-efficacy, cloud platform usage, readiness, and effectiveness of participation in online classes through Google Classroom and Zoom cloud platforms. We have not focused on a specific predictive model; instead, we have considered four machine learning algorithms to identify best model.

In the light of the reviewed literature above, cloud computing adoption in higher education institutions have a positive effect. Rest part of paper was organized as follows. Section 3 explains research method that provides a description on real time student dataset, preprocessing, feature extraction and normalization. Section 4 describes machine learning algorithms; Section 5 reveals experimental results. Section 6 draws conclusions and the further research guidelines.

### III. RESEARCH METHODOLOGY

#### A. Research Questions of Inquiry

The intention of this study is to extend an early detection model for finding at-risk students. Machine learning algorithms are implemented on the student's real time dataset collected from various educational institutions, and its accuracy has been analyzed. Within an online learning environment, the system can be globalized to any type of course. Proposed model of research is depicted in Fig. 1. This research aims to respond subsequent questions:

- 1) To build a predictive system to classify the at-risk students of cloud virtual learning platform.
- 2) To estimate model accuracy by means of various machine learning algorithms namely KNN, SVM, LDA and Random Forest.
- 3) To investigate most suited machine learning algorithm for predicting student difficulties based on demographics,

device type, access, connectivity, self-efficacy, cloud platform usage, readiness, and effectiveness of participation in online learning sessions.

#### B. Methodology

A sequence of steps has been adopted prior to the model is organized for assessment and final reporting. Methodological data flow of the complete process is visualized in Fig. 2. Initial step is collecting real data from students, and the data is preprocessed to eliminate missing value, duplications, and outliers. Feature extraction is employed to extract features. Data is normalized and is passed to the four classifiers KNN, SVM, LDA and Random Forest. Classification models have discrete outcome, we need a metric that compares discrete classes in some form. A model's performance can be evaluated using classification metrics and it determines how well or bad the classification is, but each of them evaluates it in a different way. Each classification result is tested with the performance metrics. Finally, the results are visualized.

#### C. Dataset Description

To meet the objective of the study, an electronic survey was conducted to gather real time data from higher education students belonging to various academic institutions throughout India. The instrument used for this research is E-questionnaire. It was chosen as it was considered as an efficient and effective approach. The target population for this study consisted of 530 students enrolled in online courses. The survey was broadly classified into five factors. (1) Demographic characteristics (2) Device type, mode of access, connectivity (3) Self-Efficacy (4) Familiarity with cloud platform usage (5) Readiness and Effectiveness of online learning compared to regular classroom setting. Demographic information of the respondents is presented in Table I, which shows that 66.60% are male respondents, and 33.40% are females between age group of 17 and 45 years. They are from diverse departments namely B.Sc Computer Science (20.38%), B.Com (14.34%), M.C.A (11.32%), B.C.A (11.32%), B.Tech (11.32%), M.sc (15.09%), M.com (4.91%) and BBA (11.32%) degree programmes. Online classes were managed for the students through cloud platforms. We have used free cloud services Google Forms and Google Sheets for collecting and analyzing the data. Conceptually, Self-Efficacy was estimated with 5-point Likert-scale ranging from 1= "Strongly Disagree" to 5= "Strongly Agree". Cloud platform usages are measured using 4-point range from "Not Familiar" to "Very Much Familiar". Readiness is given 4-point range from 1= "Not Ready" to 4= "Very Much Ready" and Effectiveness from 1="Much Less Effective" to 5="Much More Effective".

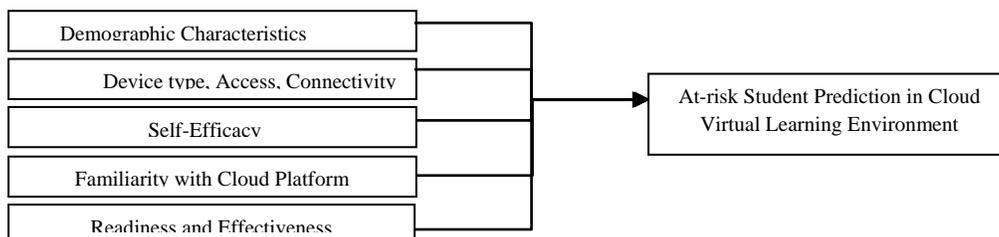


Fig. 1. Proposed Model.

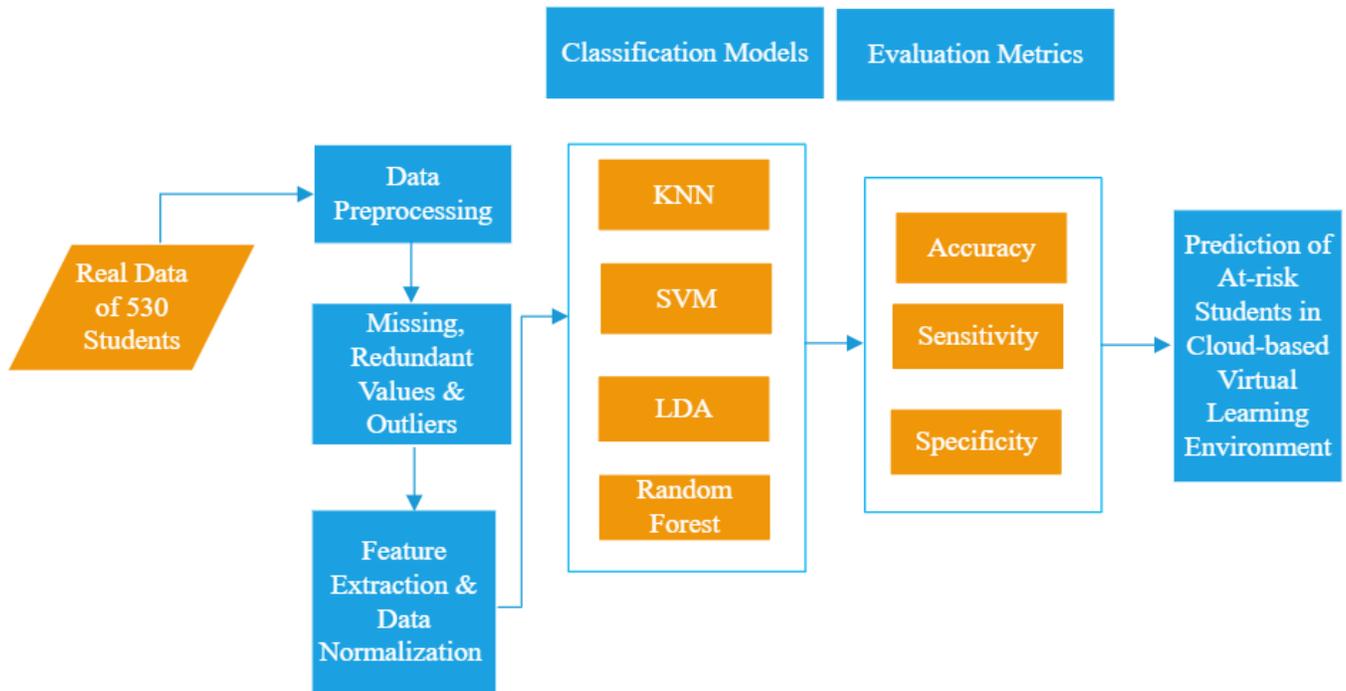


Fig. 2. Conceptual Framework of the Research Model.

TABLE I. SUMMARY OF DEMOGRAPHIC CHARACTERISTICS OF STUDENTS

| Variable                | Label                 | n=530 | %     |
|-------------------------|-----------------------|-------|-------|
| Gender                  | Male                  | 353   | 66.60 |
|                         | Female                | 177   | 33.40 |
| Age                     | 17-19                 | 158   | 29.81 |
|                         | 20-29                 | 370   | 69.81 |
|                         | 30-45                 | 2     | 0.38  |
| Education qualification | B.Sc Computer Science | 108   | 20.38 |
|                         | B.Com                 | 76    | 14.34 |
|                         | M.C.A                 | 60    | 11.32 |
|                         | B.C.A                 | 60    | 11.32 |
|                         | B.Tech                | 60    | 11.32 |
|                         | M.Sc                  | 80    | 15.09 |
|                         | M.Com                 | 26    | 4.91  |
|                         | B.B.A                 | 60    | 11.32 |

#### D. Data Preprocessing

Before evaluation phase of the classification model, the dataset was passed through a preprocessing stage. Initially, we observed that the real time dataset contains fifty-six features with missing data, redundant data, and outliers. Few students have submitted the questionnaire redundantly and missed to fill some data. When a specific value could not be estimated for a data sample, it leads to missing data. They have not provided the correct data that tends to form outliers, as the data plunge outside the range of typical distribution. By preprocessing the data, they were eliminated to maintain the quality of prediction. R4.1.1 statistical software was used as

an experimental tool. Machine learning was conducted using caret package that is a comprehensive framework for constructing machine learning models in R. The model was fed with more students' engagement activities; self-efficacy in online learning tends to learn about the behavioral patterns. The final dataset was exported into a .csv file and it is ready to be trained with different machine learning algorithms and evaluated to select most accurate one. We have used Microsoft Excel for visualizing the outcomes.

#### E. Feature Selection

It is the procedure of choosing optimal number of attributes from a larger dataset, which is the most difficult and challenging task. From this procedure, we come to know the utmost useful set of features for predicting target variable. To diminish computational cost and to increase model performance it is desirable to reduce number of features. To find the top variables, we are attaining improved cross-validated accuracy and data can be used to identify most likely elements to predict at-risk students. Two reliable measures of random forest algorithm namely %IncMSE and IncNodePurity were utilized for feature selection to generate an optimal subset of features. Finally, we have extracted forty-six features for each student with CGPA as the target variable. The selected features by our model render excellent instructions that educators can utilize to offer early assessments to learners prior to closure of course work.

1) %IncMSE: The first measure is %IncMSE which is a highly informative measure of importance. It defines the mean decrease in accuracy or how the prediction gets poorer when the variable alters its value. The variation among original

mean error and randomly permuted mean error is computed and this forms the fundamental idea for measuring important score of variables. For each tree prediction error on test is recorded (Mean Squared Error- MSE). Same process is carried after permuting each predictor variable. Higher the difference, then variable is more important. The general equation is,

$$\text{MSE} = \text{mean} ((\text{actual}_y - \text{predicted}_y)^2)$$

2) *Mean decrease gini (IncNodePurity)*: Depending on the gini impurity index this is a variable importance measure for computing splits in trees. Higher value of mean decrease gini score, higher importance of variable.

#### F. Normalization of Data

3) For machine learning this is a procedure employed as a part of data preparation. Extracted features were originally at various scales. To adjust scales of features to have a standard scale of measure the data was normalized to improve the model accuracy. The formula is given by,

$$\text{Min-Max Normalization: } (x - \min(x)) / (\max(x) - \min(x))$$

### IV. MACHINE LEARNING MODELS

Machine learning is regarding designing algorithms that automatically bring out valuable information from the given data. It is not possible for an “AI” to be trained without data. In every project, classifying and labeling datasets takes most of the time, when it reflects the real time data. The techniques applied to predict the students at risk occur as training and testing phases. To train an algorithm to know how to pertain the concepts, to identify and produce outcomes the training dataset is utilized. When we train a model on dataset, measuring its performance on the same tells how good it is at making predictions on data it has already seen. Training a model on a subset of our data, we can then use the data the model was not trained on to calculate how this would perform on unseen data. These purposefully hold out part of our dataset from training and then use the performance on this held-out dataset as a proxy of our model’s performance in production. The model is constructed on training set and examined on held-out testing set. This allows us to test that our model can generalize to unseen data. 530 students’ academic performance data were randomly divided into two datasets. Training set makes up most total data, around 70% (372 student records). The test data represents 30% (158 student records) that is used to estimate how well our algorithm was qualified with training data. Each technique is presented with data that have not been used during training to observe the classification performance during the testing phase. Before applying classifiers, feature selection methods were implemented using random forest algorithm. CGPA (Cumulative Grade Point Average) indicates the percentage of marks scored by the students. This attribute was used as response or target variable, which logically is the best predictor for course grade. This dependent variable depends on various independent variables namely demographics, previous academic outcomes, learning behaviors, device usage related factors, familiarity of cloud platform usage, self-efficacy, and readiness & effectiveness in participation of

online classes. Based on the CGPA we have calculated the risk category of students. The criteria if  $\text{CGPA} \leq 55$ , classifies the students as “At-Risk”. Else they are considered as “Non-AtRisk” students. We have used the variable Academic Performance as Boolean attribute depending upon CGPA. This attribute indicates class for supervised binary classification task and result of prediction. For our model, a negative outcome means student was not at-risk. A positive outcome defines student was at-risk. Thus, the students are classified into two categories.

For our research we have used supervised machine learning algorithms, as our dataset have labels. We considered K Nearest Neighbor (KNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA) and Random Forest (RF) as they are the most preferable algorithms by researchers to resolve related issues. We validated our methodology by supplying appropriate set of estimation measures. We have assessed the performance of different classifiers with model parameters.

#### A. KNN Algorithm

KNN categorizes a new data point into target class, based on similarity of its neighboring data points. We input a dataset of atrisk and non-atrisk students. We train our model to identify students’ performance depending on extracted features.

Choose the number k of neighbors

Compute Euclidian Distance between the data points. It is given by,

$$\text{Euclidian Distance} = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$$

Select K nearest neighbor.

Count number of data points in each group, among these K neighbors.

Allot new data point to that group for which number of neighbor is maximum.

#### B. SVM

A Support Vector Machine is utilized for classification and regression problems. In our dataset, we have used kernel SVM to conduct non-linear partitioning. The main idea behind this algorithm is,

Find the lines that separate the classes optimally. This dividing line is called a hyperplane.

Find the optimal hyperplane that helps in maximizing the margin between two classes.

Transform it into a higher dimension by employing a kernel function to dataset.

Clearly separate the two groups with a plane with the end goal of maximizing the margin.

Once the data points are separated into dimensions, SVM classifies the two groups.

#### C. LDA

Linear Discriminant Analysis (LDA) considers a data set of observations as input. We require having a categorical

variable to define class and several predictor variables for each observation. Steps involved in this algorithm,

- Compute mean vectors of each class of dependent variable
- Compute with-in class and between-class scatter matrices
- Calculate eigenvalues and eigenvector for scatter matrix within class and between class
- Sort eigenvalues in descending order and select top k.
- Create a new matrix containing eigenvectors that maps to k eigenvalues.
- Obtain linear discriminants by taking the dot product of data and matrix.

#### D. Random Forest

This model extends and integrates multiple decision trees to create a forest. It allows for more correct and constant outcomes as it relies on multitude of trees. Steps involved in this algorithm,

From the dataset having k number of records, n records are taken randomly.

For each sample individual decision trees are constructed.

Each decision tree will generate a result.

Outcome is measured depending on majority of the votes for classification or averaging the output of all trees for regression respectively.

### V. RESULT AND DISCUSSION

In this part of the study, we predicted the at-risk students from their multidimensional characteristics. To answer the research questions, we performed several experiments. The educators can utilize the predictive model to determine students having difficulties. They can deliver relevant materials, increase the student engagement activities, and improve the marks of such candidates in cloud-based learning platforms. They can acquire corrective actions at former stage, to offer supplementary assist to the students at-risk. This is vital to exactly rank the classifiers depending on their prediction potential of at-risk and subsequent decision making.

#### A. Binary Classifier Evaluation Metrics

Model performance in classification problem is assessed through confusion matrix. It represents four numbers in a two-by-two matrix. Each element displays class-wise accuracy. Reason for depicting this is to obtain benefit of our outstanding visual abilities to process more information. The elements of the confusion matrix are used to find three important parameters namely accuracy, sensitivity, and specificity. We implemented our experiment with 10-fold cross validation; each model we constructed has ensued in Kappa and Accuracy. These assessment metrics can be utilized to assess the value of classifiers for ranking various models. This generates True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Accuracy, specificity, sensitivity is used to enhance experimental results in case of binary classification.

- Accuracy is the ratio between accurate predictions over entire number of instances. It is used to determine number of times classifier is correct. This is given by,

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Total}$$

- Sensitivity (True Positive Rate) refers to proportion of correct positives that are accurately detected as positives by classifier. Formula is,

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

- Specificity (True Negative Rate) relates to classifier's ability to find negative results. The equation is,

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

Table II presents results of KNN, SVM, LDA and RF for real-time dataset. The similar data was passed for all these techniques. As the research aims to classify students' at-risk, it is vital to attain better predictive accurateness for unsuccessful learners. Accuracy from Random Forest is 88.61% it is representing good performance based on accuracy, sensitivity and specificity compared to KNN, SVM and LDA. Fig. 3 visualizes the classification accuracy, sensitivity and specificity obtained using various classifiers. The output shows our model accuracy for test set. The figure clearly depicts that Random Forest algorithm outperformed the other machine learning techniques KNN, SVM and LDA for identifying at-risk students.

TABLE II. EXPERIMENTAL RESULTS OF REAL-TIME DATASET FOR PREDICTING STUDENT'S AT-RISK USING VARIOUS CLASSIFIERS

| Classifier | Accuracy | Sensitivity | Specificity |
|------------|----------|-------------|-------------|
| KNN        | 79.11    | 97.66       | 31.32       |
| SVM        | 84.18    | 96.09       | 33.33       |
| LDA        | 82.28    | 92.97       | 36.67       |
| RF         | 88.61    | 96.09       | 56.67       |

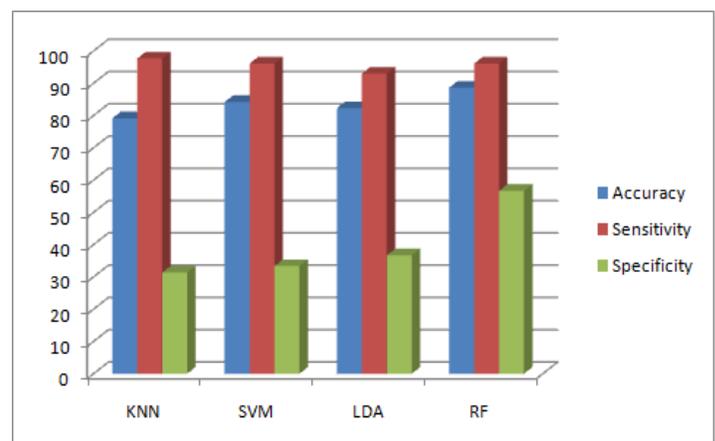


Fig. 3. Visualization Results of Different Classifier for Real-Time Student Dataset.

### B. ROC Curves

When we are evaluating classifier performance it is the dominant visualization tool. Performance metrics helps generate an aggregate perspective of a model's performance. For binary classification problems, receiver operating characteristic (ROC) curves can be very informative. It illustrates true positive rate as a function of false positive rate, hence prompting sensitivity of classifier. We partition predictions into positive and negative classes for purpose of obtaining ROC measurements namely specificity and sensitivity normally used on ROC curve axis. For our real-world dataset, if we are classifying whether a student is at-risk or non-atrisk, it is significantly better to categorize a small number of additional non-atrisk students as at-risk and avoid classifying any at-risk as non-at-risk students. We would choose threshold that reduces false-negative rate, increase true-positive rate, and rest us at the top of the ROC plot. This gives an observation of overall performance of the classifier. Fig. 4 visualizes column-wise area under ROC curve (AUC) for KNN, SVM, LDA and RF. For classification models, another performance metric is AUC. This is measure of capability of classifier to differentiate among classes and is

utilized as a summary of ROC curve. The classification performance is enhanced when this area is larger. For evaluating and comparing models this is an extensively used option.

An ROC curve gives us a more nuanced view of how a model's performance changes as we make predictions conservative. Table III illustrates the results of ROC curve for various machine learning algorithms. After observing ROC curves of each classifier, it's clear that Random Forest algorithm has the highest statistic of ROC 76.38%. Classification models often use the area under the curve (AUC) to represent performance. It delivers an agreement measure of performance across all possible classification thresholds. Fig. 5 illustrates the highest Area Under Curve that corresponds to random forest algorithm. Higher AUC depicts classifier has outstanding performance to differentiate among positive and negative classes. From these series of experiments, it is apparent that in overall random forest algorithm can be used to detect the students regarded at-risk in cloud-based virtual learning environment based on multidimensional variants.

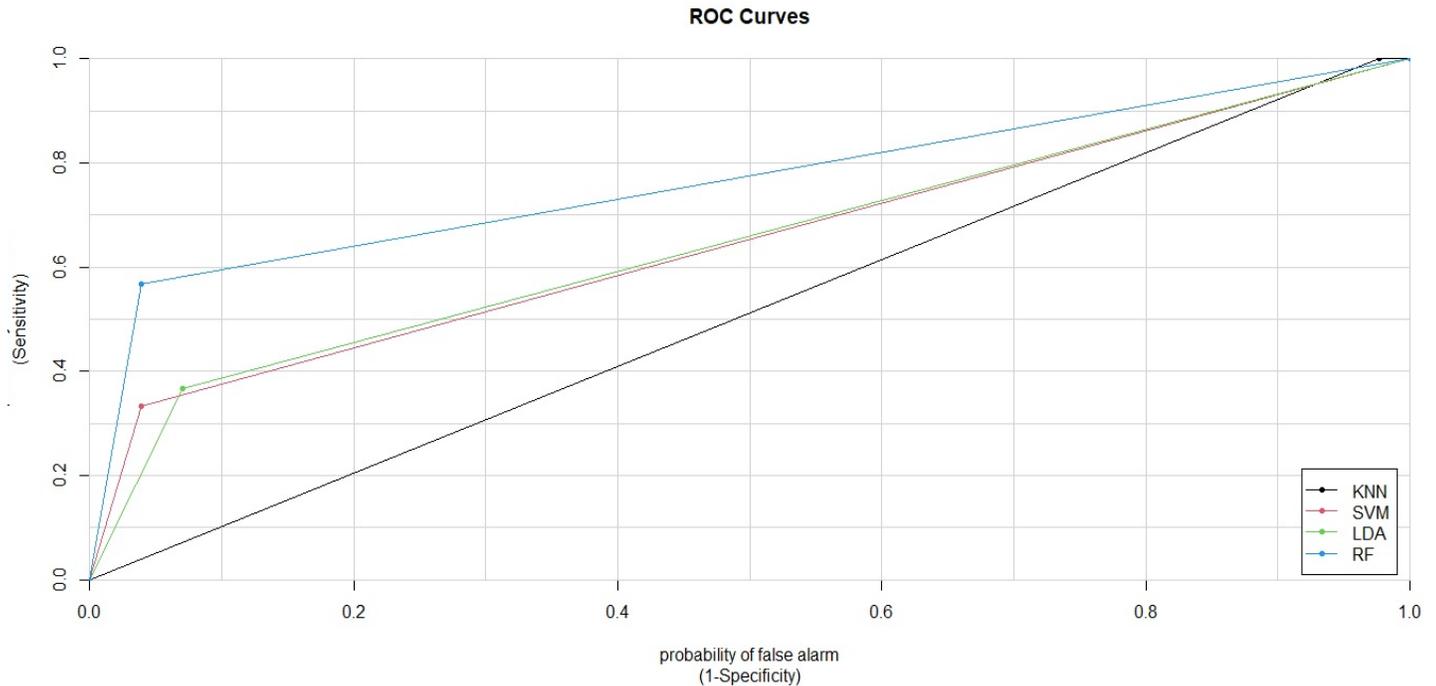


Fig. 4. Column-Wise Area Under ROC Curve (AUC) of Classifiers for Students' At-Risk Prediction.

TABLE III. COMPARATIVE RESULTS OF RECEIVER OPERATOR CHARACTERISTIC (ROC) CURVES FOR KNN, SVM, LDA AND RF

| Classifier | ROC Curve Statistic |
|------------|---------------------|
| KNN        | 51.17               |
| SVM        | 64.71               |
| LDA        | 64.82               |
| RF         | 76.38               |

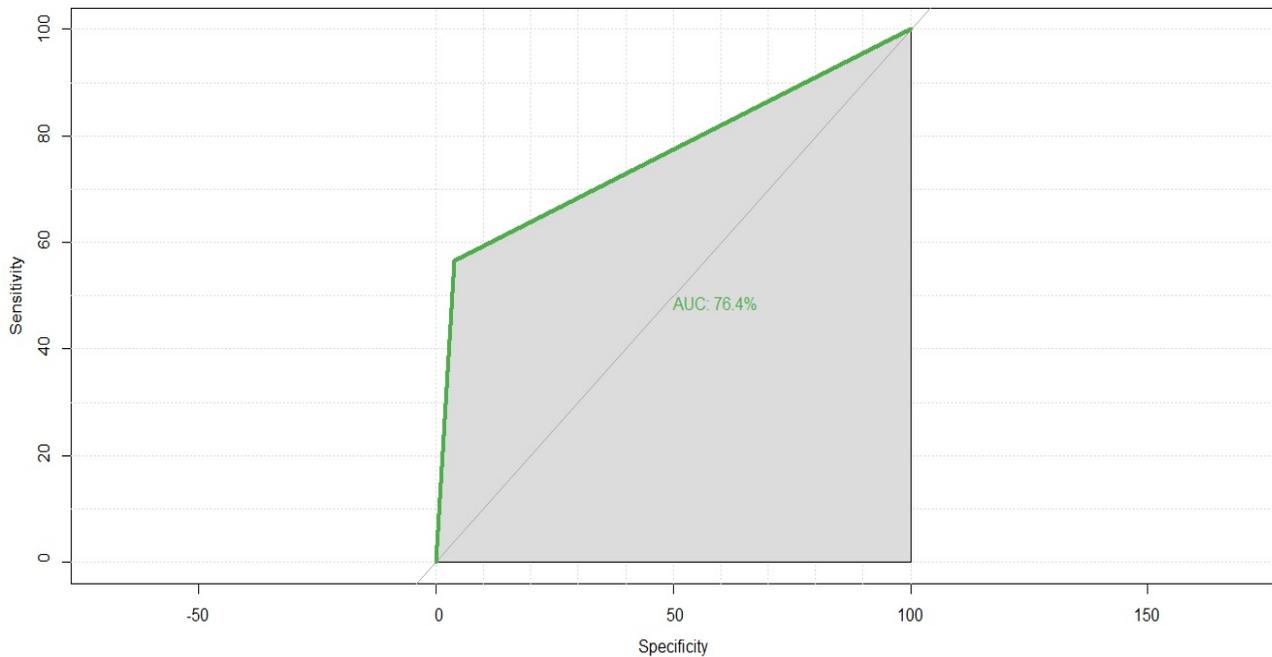


Fig. 5. ROC Curve Illustrating the Highest Area Under Curve corresponding to Random Forest Algorithm.

## VI. CONCLUSION AND FUTURE WORK

As there is a sudden shift away from the traditional classroom in many parts of the globe, the adoption of virtual learning is more effective during the COVID-19 crisis. The students are in diverse geographical locations, the cloud platforms have empowered them to learn in their own comfort zone. They had an opportunity to participate in an interactive and collaborative learning environment. The instructors can upload and share the course material in cloud platforms. The students can access the materials from their own individual portal that enables them as independent learners. By applying machine learning algorithms in cloud platforms progress of the students can be tracked hence the learning gaps can be identified. This research work focuses on implementing students' outcome based early prediction model. We employed different machine learning approaches namely KNN, SVM, LDA and RF to detect at-risk students. The lower academic results decrease self-confidence of the learners and depletion of valuable educating efforts. These students always have an intention of dropping out from college. To solve the problem of drop out, this system helps the higher education institutions to classify the risky students at earlier stage. The amount of data generations depends on the students' multivariate characteristics for the courses enrolled in online classes. This was used to create the machine learning model that resourcefully utilizes this data, hence forth bring outcomes that can be used additional in students' wellbeing in terms of their performance and personal growth. The versatility of these systems also helps the teachers to take potential efforts towards the risky students. A sequence of experiments has been managed to identify the most excellent model.

Comparing to the existing research and results, current research revealed that the most promising random forest

algorithm achieved high accuracy with 88.61% and outperformed other binary classification models. These algorithms were used to classify the students' at-risk in VLE by considering their multideterminant characteristics. The outcomes from this model can profoundly help educators, to upgrade their existing teaching methodologies and the implementation of new techniques facilitates the students to pay additional attention in their studies. This data-driven study can support VLE administrators, instructors, and course co-coordinators in the articulation of effective virtual learning structure that can bestow to process of decision-making. This early intervention technique that was implemented in virtual learning environment motivates the students to have high academic scores.

Depending on our results machine learning is highly recommended to be integrated with cloud platforms. It gives an insight into real-time situations that allows the higher educational institutions to forecast future outcomes. In further research, we plan to deploy predictive model in cloud computing platform by means of helping the educators to progress the performance of unprepared students and for automating the decision-making process.

## REFERENCES

- [1] Z. M. A. Abdullah Alghushami, Nur Haryani Zakaria, "The determinants impacting the adoption of cloud computing in Yemen institutions," in *AIP Conference Proceedings*, 2018, pp. 1–7, Available: <https://doi.org/10.1063/1.5055424>.
- [2] H. C. & W. Z. Ahmed A. Mubarak, "Prediction of students' early dropout based on their interaction logs in online learning environment," *Interact. Learn. Environ.*, 2020.
- [3] D. A. A. and R. Masa'deh, "Antecedents of students' perceptions of online learning through covid-19 pandemic in Jordan," *Int. J. Data Netw. Sci.*, vol. 5, no. 4, pp. 587–592, 2021.
- [4] M. S. David Baneres, M.Elena Rodriguez-Gonzalez, "An Early Feedback Prediction System for Learners At-Risk Within a First-Year

- Higher Education Course,” *IEEE Trans. Learn. Technol.*, vol. 12, no. 2, pp. 1–14, 2019.
- [5] D. Fong, “Redesigning Prediction Algorithms for At-Risk Students in Higher Education: The Opportunities and Challenges of Using Classification Techniques in a University Academic Writing”, Book: “Redesigning Higher Education Initiatives for Industry 4.0,” IGI Global, pp. 232-250, 2019.
- [6] Y. S. Edward Wakelam, Amanda Jefferies, Neil Davey, “The potential for student performance prediction in small cohorts with minimal available attributes,” *Br. J. Educ. Technol.*, vol. 51, no. 2, pp. 347–370, 2020.
- [7] E. García-Salirrosas, “Satisfaction of university students in virtual education in a COVID-19 scenario,” 3rd International Conference on Education Technology Management (ICETM), 2020, pp. 41-47.
- [8] S. P. Z. Francesca Del Bonifro, Maurizio Gabbrielli, Giuseppe Lisanti, “Student Dropout Prediction,” *21st Int. Conf. Artif. Intell. Educ. AIED*, pp. 129–140, 2020.
- [9] D. R. G. Francis Ofori, Dr. Elizaphan Maina, “Using Machine Learning Algorithms to Predict Students’ Performance and Improve Learning Outcome: A Literature Based Review,” *J. Inf. Technol.*, vol. 4, no. 1, pp. 33–55, 2020.
- [10] H. O. GokhanAkcinar, Mohammad Nehal Hasnine, Rwitajit Majumdar, Brendan Flanagan, “Developing an early-warning system for spotting at-risk students by using eBook interaction logs,” *Smart Learn. Environ.*, vol. 6, no. 4, pp. 1–15, 2019.
- [11] A. F. Hassan Zeineddine, Udo Braendle, “Enhancing prediction of student success: Automated machine learning approach,” *Comput. Electr. Eng.*, vol. 89, no. 4, 2020, [Online]. Available: <https://doi.org/10.1016/j.compeleceng.2020.106903>.
- [12] A. F. Ihsana El Khuluqo, Abdul Rahman A. Ghani, “Postgraduate students’ perspective on supporting ‘learning from home’ to solve the COVID-19 pandemic,” *Int. J. Eval. Res. Educ.*, vol. 10, no. 2, pp. 615–623, 2021.
- [13] R. G.-C. Ivan Sandoval-Palis, David Naranjo, Jack Vidal, “Early Dropout Prediction Model: A Case Study of University Leveling Course Students,” *Sustainability*, vol. 12, no. 9314, pp. 1–17, 2020.
- [14] J. B. Johannes Berens, Kerstin Schneider, Simon Gortz, Simon Oster, “Early Detection of Students at Risk- Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Models,” *J. Educ. Data Min.*, vol. 11, no. 3, pp. 1–41, 2019.
- [15] K. Alhumaid, “Developing an educational framework for using mobile learning during the era of COVID-19,” *Int. J. Data Netw. Sci.*, vol. 5, no. 3, pp. 215–230, 2021.
- [16] T. M. L. Kwok Tai Chui, Dennis Chun Lok Fung, Miltiadis D. Lytras, “Predicting At-risk University Students in a Virtual Learning Environment via a Machine Learning Algorithm,” *Comput. Human Behav.*, vol. 107, 2020, Available: <https://doi.org/10.1016/j.chb.2018.06.032>.
- [17] N. A. Leena H. Alamri, Ranim S. Almuslim, Mona S. Alotibi, Dana K. Alkadi, Irfan Ullah Khan, “Predicting Student Academic Performance using Support Vector Machine and Random Forest,” in *3rd International Conference on Education Technology Management (ICETM)*, 2020, pp. 1–8.
- [18] M. Revani Putri, K. Oktriono, C. Sidupa, M. Willyarto “Portraying Students’ Challenges and Expectations toward Online Learning in Embracing Industrial Revolution 4.0 Era: A case in ELT in the COVID-19 Outbreak,” 3rd International Conference on Education Technology Management (ICETM), 2020, pp. 36-40.
- [19] M. B. and S. U. K. Muhammad Adnan, Asad Habib, Jawad Ashraf, ShafaqMussadiq, Arsalan Ali Raza, Muhammad Abid, “Predicting at-Risk Students at Different Percentages of Course Length for Early Intervention Using Machine Learning Models,” *IEEE Access*, vol. 9, pp. 7519–7539, 2021.
- [20] S. M. R. A. Mushtaq Hussain, Wenhao Zhu, Wu Zhang, “Student Engagement Predictions in an e-Learning System and Their Impact on Student Course Assessment Scores,” *Comput. Intell. Neurosci.*, vol. 2018, no. 6347186, pp. 1–21, 2018.
- [21] S. A. Mushtaq Hussain, Wenhao Zhu, Wu Zhang, Syed Muhammad Raza Abidi, “Using machine learning to predict student difficulties from learning session data,” *Artif. Intell. Rev.*, vol. 52, pp. 381–407, 2019, Available: <https://doi.org/10.1155/2018/6347186>.
- [22] S.-U. H. Naif RadiAljohani, Ayman Fayoumi, “Predicting At-Risk Students Using Clickstream Data in the Virtual Learning Environment,” *Sustainability*, vol. 11, no. 24, pp. 1–12, 2019.
- [23] P. NorkaBedregal-Alpaca, Víctor Cornejo-Aparicio, Joshua Zárate-Valderrama, “Classification Models for Determining Types of Academic Risk and Predicting Dropout in University Students,” *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 1, pp. 266-272, 2020.
- [24] K. S. and V. S. Shirin Riazy, “Fairness in Learning Analytics: Student At-risk Prediction in Virtual Learning Environments,” *In Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*, 2020, pp. 15–25.
- [25] F. M. Z. Zulherman, ZalikNuryana, AstadiPangarso, “Factor of Zoom cloud meetings: Technology adoption in the pandemic of COVID-19,” *Int. J. Eval. Res. Educ.*, vol. 10, no. 3, pp. 816–825, 2021.
- [26] M. D. Janka Kabathova, “Towards Predicting Student’s Dropout in University Courses Using Different Machine Learning Techniques,” *Appl. Sci.*, vol. 11, no. 7, pp. 1–19, 2021, Available: <https://doi.org/10.3390/app11073130>.
- [27] J. A. C. R. Luis Earving Lee, Salvador Ibarra Martinez, M. G. T. B. Jesus David Teran Villanueva, Julio Laria Menchaca, and E. C. Rocha, “Evaluation of Prediction Algorithms in the Student Dropout Problem,” *J. Comput. Commun.*, vol. 8, pp. 20–27, 2020, Available: <https://doi.org/10.4236/jcc.2020.83002>.
- [28] S. J. Y. Hee Sun Park, “Early Dropout Prediction in Online Learning of University using Machine Learning,” *Int. J. Informatics Vis.*, vol. 5, no. 4, pp. 347–353, 2021, Available: <http://dx.doi.org/10.30630/joiv.5.4.732>.

# Balanced Schedule on Storm for Performance Enhancement

Arwa Z. Selim<sup>1</sup>, I. M. Hanafy<sup>3</sup>

Department of Mathematics and Computer Science  
Faculty of Science, Port Said University  
Port Said, Egypt

Noha E. El-Attar<sup>2</sup>, Wael A. Awad<sup>4</sup>

Faculty of Computers and Artificial Intelligence<sup>2,4</sup>  
Benha University, Benha, Egypt<sup>2</sup>  
Damietta University, Damietta, Egypt<sup>4</sup>

**Abstract**—In recent years, real-time and big data aroused and received a lot of attention due to the spread of embedded systems in almost everything in life. This has led to many challenges that need to be solved to enhance and improve systems that work on big real-time data. Apache Storm is a system used for computing and analyzing big real-time data of distributed systems. This paper aims to develop a scheduler to improve the scheduling of the applications represented by topologies on the Storm cluster. The proposed scheduler is hybridization between the scheduling algorithms of A3 Storm and the Workload scheduler. Its objective is to minimize the communication between tasks while balancing the workload on all cluster machines. The proposed scheduler is compared with the A3 Storm and Fischer and Bernstein's scheduling algorithm. The comparison has been made using four different topologies. The experimental results show that our proposed scheduler outperforms the two other schedulers in throughput and complete latency.

**Keywords**—Real-time; big data; apache storm; scheduling

## I. INTRODUCTION

Real-time applications such as IoT sensors, climate, and healthcare produce a large amount of continuous real-time data. The nature of this type of data is overgrowing where it can reach quintillions of bytes every day. This extreme and rapid growth of data leads to the term "big data" [1]. 5Vs features usually characterize big data; volume, variety, velocity, veracity, and value. Volume refers to the massive amount of data [2]. Variety means that there are different types of data that cause complexity. The rate at which the data is produced and transferred is the velocity, and it must be analyzed in real-time. Veracity is the precision level of data. Finally, the value is the valuable information derived from the data [1] [3]. Generally, the "big data" processing can be done through two processing techniques; batch and stream processing on high-performance computing resources [4]. Batch processing works on data that is previously stored. At the same time, stream processing refers to processing a large amount of data in real time. Big Data needs specified applications for processing the data, such as Hadoop for batch processing and Apache Storm, S4, Spark, and Flink for real-time streaming applications [4].

Real-time refers to the concept of time quantity, which implies the necessity for a real-time clock to measure it [5]. The real-time tasks are classified into three different cases: hard real-time, firm real-time, and soft real-time. The constraint in hard real-time tasks is to create results within

specific time constraints or cause disastrous results. Firm real-time tasks also have to create the results before the specified time constraints, or the results will be invaluable. Soft real-time tasks have no time limitation, the results could be generated at any time, and it will be beneficial and acceptable [5] [6].

Now-a-days, the processing of streaming data is gaining more attention due to its sensitive cases. Thus, many researchers have tried to enhance the scheduling techniques to handle this huge amount of data and increase performance of processing (i.e., decreases the latency, increase the throughput, balance the network load, etc.). Most of the researchers' algorithm achieved those performance objectives. The relevant algorithms some of them achieved increment in the throughput, some achieved network load enhancement, others achieved decrement in latency, etc. Real-time scheduling for streaming data needs continuous improvements to get better results with better performance. So the issue here is to propose an algorithm that increases the performance of the processing of streaming data in using Apache Storm.

The main idea of the proposed algorithm is to reduce the communication between executors. The scheduler collects information during runtime then it creates a schedule using graph partitioning technique that partition the communication graph [7] [8]. At the end, the collected communication between executors during runtime is used and the pairs of most communicating executors are assigned to the same slot. This will achieve workload balance; improve resource utilization, high throughput with more reduced load on network.

The paper contribution is as follows:

- 1) We proposed a hybrid between two algorithms, the Workload scheduling algorithm and the A3 Storm algorithm, which improves the performance of the Apache Storm. This scheduler maximizes the throughput and minimizes the latency as the communication network load is reduced.
- 2) The scheduler is based on graph partitioning; its objective is workload balance and inter-executor communication.
- 3) Four topologies evaluate the scheduler, and the metrics of throughput and latency are collected as results.
- 4) A comparison is made against two alternative schedulers to find which has better results when running the topologies on them.

This paper is organized as follows. In Section 2, we introduce an overview of the Storm and its scheduling. The related work is shown in Section 3. Section 4 discusses the state-of-the-art scheduling algorithms. The proposed algorithm is discussed in Section 5. Section 6 discusses the results with comparison to the state-of-the-art algorithms. Section 7 is the conclusion.

## II. OVERVIEW ON APACHE STORM

Apache Storm is a widely used real-time processing framework due to its capabilities of carrying out analytics on data streams with high throughput. Storm is an open-source real-time processing framework that contains several components: Topology, Nimbus, Slave, and zookeeper. [9]. Topology is considered the main component in Storm; it is a directed acyclic graph (DAG) that consists of spout and bolts [9] [10]. The spout is the primary source of a stream, and the bolt is the processing unit of the topology which handles the stream of data. Storm is deployed on a cluster that follows a master-slave model. The master is called the Nimbus node, which is responsible for organizing the topology and analyzing it [11][10]. It also distributes the tasks among the supervisor nodes and monitors any failure occurrence on them (i.e., if one of the supervisors fails, it redistributes the work among the remaining supervisors). The slave is any supervisor node. Storm can have one or more supervisors, and each supervisor can have one or more workers, which helps execute the tasks assigned by the supervisor. The worker also can have one or more executors which are responsible for running and executing the tasks. In the end, the task carries out the actual processing of the data (i.e., the tasks can be spouts or bolts). Also, Storm contains the zookeeper, which coordinates the work between the Nimbus and supervisors and saves their state (i.e., in case of Nimbus failure, it will restart from its last state as it is saved in zookeeper). Fig. 1 depicts the components of the Storm cluster.

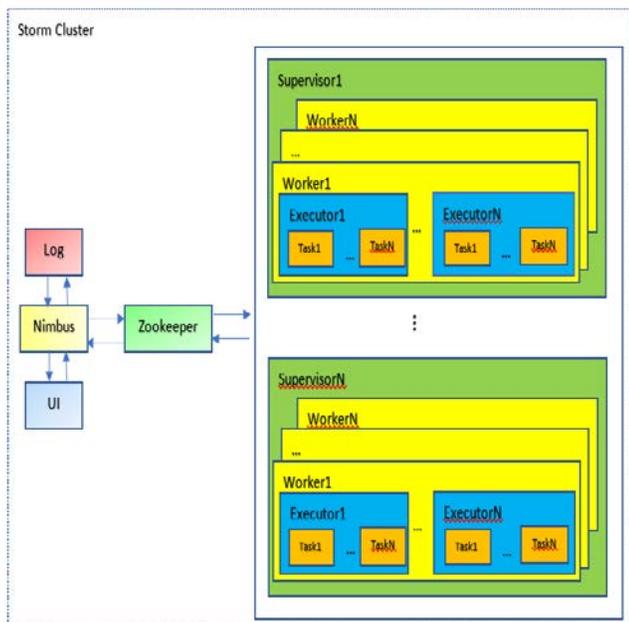


Fig. 1. Storm Cluster.

A default scheduler performs the standard scheduling process in Storm existed in Nimbus called EvenScheduler. This scheduler goes through two main phases; the first is assigning the executors to the workers, and the second is allocating the workers to the slots. EvenScheduler works based on round-robin strategy as follows [3]:

It iterates through the executors of the topology and allocates every executor evenly to the workers based on a round-robin algorithm.

The workers are allocated and assigned evenly to the supervisors, considering the available slots in each supervisor.

## III. RELATED WORK

One of the main benefits of Storm is that it is an open-source framework that allows creating custom schedulers that can meet the needs of the users and data. The default scheduler in Storm has some drawbacks which need to be improved. For instance, it evenly assigns the tasks to the cluster slots, but it does not consider the inter-node and inter-slot communication. Thus, the traffic may negatively influence the throughput and performance of the processing. Recently, many researchers have proposed enhanced scheduling algorithms that can improve the performance of Storm.

Aniello L. et al. [3] have proposed two scheduling algorithms, offline and online. The offline scheduling algorithm is based on the topology structure and how its components are interconnected with each other. It used the round-robin algorithm to assign slots to nodes. The online scheduling algorithm was based on monitoring the communication and the performance of the system at run time. It monitored the traffic of exchanged tuples between executors, sorted the executors in descending order according to the communication patterns, and assigned the most communicating executors in the same slot. Then the communication pairs of workers are iterated in descending order and assigned to the nodes. This algorithm reduced the inter-node traffic and communication, which affected the network load and the throughput.

Xu J. et al. [12] have developed a traffic-aware online scheduling algorithm that can reduce the inter-node and inter-process traffic by monitoring the traffic and workload information during runtime. It expedites the data processing by using the traffic-aware online algorithm for assigning executors. It also enables fine-grained control over worker node consolidation to obtain better performance while using fewer worker nodes.

In the same context, Peng B. et al. [11] have presented a resource-aware scheduling algorithm to improve resource utilization and reduce network latency. This algorithm is based on assigning the task according to an improved breadth-first traversal algorithm. It allocates the node ports that conform to the resource constraints and the network distance requirement.

Fischer L. and Bernstein A. [8] have proposed a workload scheduling based on the graph partitioning technique. It works during runtime and collects the behavior of communication of the topologies that are running. Then it partitions the communication graph to produce the schedule using software

called METIS graph partitioning. METIS uses a multilevel graph bisection. It makes a miniature version of the graph by coarsening it and collapsing the nodes and edges. Then it partitions the resulting small graph before un-coarsening it to its first form. It adapts the partition at each step of the un-coarsening to consider the newly un-collapsed edges and vertices. This scheduler improves resource utilization, reduces the network load, and increases the throughput.

Another direction is presented by Li C. et al. [13]. They have developed a Storm topology dynamic optimization strategy (STDO-TOC) as a real-time scheduling algorithm. The STDO-TOC uses bolts capacity and analyzes the message queue congestion degree to alter the performance parameters during runtime. If the topology bottlenecks are found, they are automatically removed. This leads to optimizing the topology dynamically.

Another resource-efficient algorithm for streaming application scheduling D-Storm has been presented in Liu X. and Buyya R. [14]. D-Storm tracks the streaming tasks during runtime to collect resources and communications and use them in the scheduling process to pack the communicating tasks compactly. This strategy of tight scheduling reduces resource utilization and inter-node communication.

Muhammad A. et al. [15] have developed a topology-based resource-aware scheduling algorithm called Top Storm. It is based on finding the most communicating executors and putting them closer to each other to reduce the number of nodes used to execute topology. Top-Storm is considered a topology-based as it looks at the DAG of the topology to find the connections between the executors. Also, it is resource-aware as the assigning of executors is made based on the computation power of nodes.

An enhanced version of Top Storm called A3 Storm has been developed by Muhammad A. and Aleem M. [4]. It works offline by finding the most communicating executors from the DAG of the topology and putting them closer to each other, and assigning them to the nodes according to the most powerful one. At the same time, it can work online by using traffic beside the topology structure. It reads the inter executor traffic, sorts it into descending order, and then assigns it to the most influential nodes. This scheduler improves resource utilization and increases the throughput of the topology. Finally, Table I displays a brief review of the above-mentioned scheduling algorithms.

TABLE I. A BRIEF REVIEW OF REAL-TIME SCHEDULING ALGORITHMS

| Scheduling Aspects             | Scheduling Algorithms Characteristics |               |         |              |               |                |               |
|--------------------------------|---------------------------------------|---------------|---------|--------------|---------------|----------------|---------------|
|                                | Resource Aware                        | Traffic-Aware | Dynamic | Heterogenous | Self-Adaptive | Topology Aware | Network-Aware |
| Aniello, et al., 2013) [3]     | x                                     | ✓             | ✓       | ✓            | ✓             | ✓              | x             |
| Xu, et al., 2014 [12]          | x                                     | ✓             | ✓       | ✓            | ✓             | x              | x             |
| Peng, et al., 2015 [11]        | ✓                                     | x             | x       | ✓            | x             | x              | ✓             |
| Fischer & Bernstein, 2015) [8] | ✓                                     | ✓             | ✓       | ✓            | ✓             | x              | ✓             |
| Li, et al., 2017 [13]          | ✓                                     | ✓             | ✓       | ✓            | ✓             | ✓              | x             |
| Liu & Buyya, 2017 [14]         | ✓                                     | ✓             | ✓       | ✓            | ✓             | x              | x             |
| Muhammad, et al., 2021 [15]    | ✓                                     | x             | x       | ✓            | x             | ✓              | x             |
| Muhammad & Aleem, 2021) [4]    | ✓                                     | ✓             | ✓       | ✓            | ✓             | ✓              | x             |

#### IV. PRELIMINARIES

##### A. Workload Scheduler

The workload scheduler is the standard storm topology. It has two types of views, logical view, and physical view. The logical view,  $T = (N, C)$ , consists of  $N$  spouts and bolts connected with  $C$  number of connections. While the physical view of storm topology is represented by graph  $G = (V, E)$ , where  $V$  represents the vertices, and  $E$  represents the edges of the graph. A set of task instances represents the spout and bolts  $v_i \in V$ , and  $|v_i| = d_i$  represents the degree of parallelism of each component, spout, or bolt. Every two sets of vertices  $V$  are connected by one edge  $E$ . Generally, the graph is weighted as follows [8]:

- The vertex weights are represented by the sum of the number of all released and received tuples.
- The edge weights are the number of messages released from any of the spout or bolt instances.

The main idea in graph partitioning is to partition the vertices into equal partitions to reduce the edges' number connecting the vertices of different partitions based on the k-way partitioning method. To clarify the k-way partitioning, if a given graph  $G = (V, E)$ , the vertices will be partitioned into  $M$  number of partitions  $P$ , where  $M$  is equal to the supervisor machines number in the cluster. Where  $\cup_{m=1}^M P_m = V$  and  $\cap_{m=1}^M P_m = \emptyset$ . [8] [16].

The communication between partitions can be represented as a matrix. If there is a task  $\tau_u$  and partition  $P_a$ . To check if the task  $\tau_u$  is assigned to the partition  $P_a$ :

$$M_{\tau_u, P_a} = \begin{cases} 1, & \text{if the task } \tau_u \text{ is assigned to the partition } P_a. \\ 0, & \text{if the task } \tau_u \text{ is not on the partition } P_a. \end{cases} \quad (1)$$

Then the communication between the nodes in the communication graph can be represented as follows in (2) [8]:

$$C_{\tau_u, \tau_v} = \sum_{a=1}^P \sum_{b=1}^P M_{\tau_u, P_a} \times M_{\tau_v, P_b} \quad (2)$$

Where  $M$  is the total number of partitions.

According to (2), the formula of the node's communications can be summarized as follows:

$$C_{\tau_u, \tau_v} = \begin{cases} 1, & \text{if task } \tau_u \text{ and task } \tau_v \text{ are on different partitions} \\ 0, & \text{if task } \tau_u \text{ and task } \tau_v \text{ are on the same partitions} \end{cases} \quad (3)$$

Finally, the cost function of the partitioned graph  $G$ , which is partitioned into  $M$  partitions, can be defined as follows [8]:

$$cost(G, M) = \sum_{u=1}^{|V|} \sum_{v=1}^{|V|} C_{\tau_u, \tau_v} \times e_{u,v} \quad (4)$$

where  $|V|$  is the total number of vertices of the graph and  $e_{u,v}$  represents the edge weights.

This partitioning algorithm aims to optimize the costs for the partitions. In this algorithm, there is another constraint: the partitions should be balanced for the workload. The load balance factor  $L_{im}$  is defined as below [8]:

$$L_{im}(G, M) = \max(P_i / AP) \quad (5)$$

where  $P_i$  is the sum of weights of all vertices in partition  $i$ , and  $AP$  is the average weight of partition over all partitions.

The workload scheduler uses METIS as software for graph partitioning, which is used for partitioning a graph into equal partitions [16]. The partitioning process includes three main phases: coarsening, partitioning, and un-coarsening.

Initially, the coarsening phase reduces the size of the graph by combining a set of vertices into one single vertex. The weight of this single vertex must be equal to the sum of weights of all vertices combined in it.

Then the partitioning phase is done on the coarsest graph to receive balanced partitions with respect to the workload.

Finally, the un-coarsening phase is used to return to the original graph and refine the resulting partitions.

To assign the partitions, their number must be equal to the number of machines in the cluster. Each partition will be assigned to one supervisor, and its tasks will be assigned to the slots. Finally, the executors in each partition are assigned to slots in a round-robin manner [7]. A flowchart of the workload scheduler is illustrated in Fig. 2.

##### B. A3 Storm Scheduler

A3 Storm scheduler is a topology, traffic, and resource-aware scheduler. It can find the connections between the executors by considering the DAG of topology, so it is considered a topology-aware scheduler. It also puts inter-executor communication into consideration, so it is a traffic-aware scheduler. Finally, it performs the physical mapping according to the computation power of nodes; thus, it also resources aware scheduler [4].

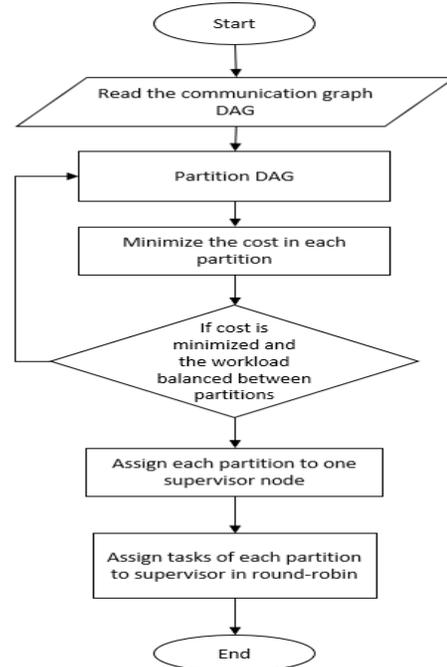


Fig. 2. Workload Scheduler Flowchart.

In general, the A3 Storm scheduler follows two steps: (1) Assignment of Executors, either by using the DAG or the traffic of topology. In this step, the scheduler initially gets the number of worker processes and the inter-executor traffic for executors unassigned from the traffic log. Then the executor assignment process begins by sorting the executors in descending order according to the inter-executor traffic; after that placing the most communicating executors as close as possible to each other. (2) Assignment of Slot; in this step, the created groups of the highest communicating executors are assigned slots. Then these slots are assigned to nodes which are sorted in descending order according to the most computationally powerful node calculated by (6) and (7). [4] [15].

$$Computation_{power} = \alpha \times (Speed) + (1 - \alpha) \times RAM \quad (6)$$

$$Speed = no. of. Cores \times no. of. Sockets \times frequency \times no. of. Flops \quad (7)$$

Where,  $\alpha$  is an adjustment factor equal 0.8.

#### V. FORMULATION OF THE BALANCE WORKLOAD STORM SCHEDULER ALGORITHM

The proposed algorithm hybridizes the Workload scheduler [8] and the A3 Storm scheduler [4]. As mentioned before, the Workload scheduler aims to reduce the network utilization by reducing the inter-node communication to increase the overall throughput and balance the workload among all available machines in the cluster. At the same time, the objective of the A3 Storm scheduler is to reduce the inter-node traffic, increase the throughput, and improve resource utilization.

The proposed algorithm of Balance Scheduling on Storm (BSS) is based on combining both the Workload scheduler and A3 Storm to enhance the performance of the Workload scheduler in increasing the data throughput and reducing the time latency. The main idea of the BSS is to apply the methodology of workload scheduler to monitor the metrics of Storm, collect the data of the communication graph during runtime, partition the graph, and balance the workload among the available machines in the cluster. Then, we have replaced the last step in the Workload scheduler (i.e., assigning each partition to one node and assigning tasks to slots in a round-robin strategy) by the A3 storm first phase (i.e., monitoring the communication between tasks). Therefore, after assigning the partitions to the nodes, the tasks will be assigned to the slots according to their inter-executor communication, which means that the most communicating executors will be assigned to the same slot, which will reduce the network communication between slots and in result this will reduce the latency and increase the throughput.

##### A. The Proposed Algorithm of Balance Scheduling on Storm (BSS)

The proposed algorithm of Balance Scheduling on Storm is based mainly on managing the execution of the topology. Its input and output are the unassigned topologies and the executor to node assignment, respectively. The phases of the proposed BSS algorithm can be concluded as follows:

1) The scheduler collects information about the topology, where it gets the number of worker processes (slots) required to execute the topology. Then it obtains the inter-executor connection of the unassigned executors and the number of the unassigned executors.

2) Calculating the maximum number of executors required per slot by (8):

$$e_{es} = \frac{total\_number\_of\_executors}{total\_number\_of\_slots\_required\_by\_topology} \quad (8)$$

3) Finding the available nodes in the cluster.

4) Partitioning the vertices of the graph into equal partitions using the k-way partitioning method, where the number of partitions should be equal to the number of the topology workers.

5) Sorting the nodes where the nodes with no workers are put at the beginning, and the partially busy nodes are put at the end. In this stage, the scheduler has to ensure that the number of nodes is not smaller than the number of partitions.

6) Sorting the tasks in each pair in descending order according to their inter-executor traffic.

7) Assigning tasks to the slots, where the most communicating tasks are assigned to the same slot until it reaches the maximum number of tasks assigned to it. Then assigning the next task to the next slot and so on until all partition tasks are assigned to the slots of the node. The pseudo-code of the proposed algorithm of Balance Scheduling on Storm is shown in Algorithm 1.

---

**Algorithm (1): Balance Scheduling on Storm**

---

```
1. function Schedule (U);
Input: Unassigned Topologies U
Output: Node-Executor assignment
2. for each topology $u_i \in U$ do
3. $n = u_i.numOfWorkers$;
// get the number of worker processes
4. $e_{un} = u_i.UnassignedExecutors()$;
// get the list of Inter-Executor for unassigned executors
5. $e_{total} = e_{un}.Count()$;
6. $e_{es} = ceil(e_{total}/n)$

7. Nodes = cluster.getAvailableNodes();
8. partition_file = graph.part(Nodes);
// partition = number of workers
9. Nodes.sort();
// Nodes that have no worker are put at the beginning and
// partially busy nodes are put at the end
10. $slots_{as} = 0$;
11. if Nodes.size() >= partition.size() do
12. Foreach $n \in Nodes$ do
13. Foreach $tasks \in partition.tasks$ do
14. tasks.sort("Desc");
// task pairs are sorted in descending order according to the
// InterExecutor traffic
15. While tasks != null
16. For each task $\in tasks$
17. If Count <= e_{es}
18. mapExecutorToSlot($slots_{as}$, task);
// Assign most communicating task pairs to the same slot;
19. Count ++;
```

---

```
20. End If
21. slotsas ++;
22. End for
23. End While
24. End for
25. End for
26. end If
27. End for
28. End
```

Also, all the utilized notations and functions in the algorithm are described in Table II.

TABLE II. LIST OF NOTATIONS

| Symbol                      | Definition                                                                            |
|-----------------------------|---------------------------------------------------------------------------------------|
| U                           | Unassigned topologies                                                                 |
| N                           | Total number of workers required for the execution of the topology                    |
| $e_{un}$                    | List of all unassigned executors of a topology                                        |
| $e_{total}$                 | Total number of topology executors                                                    |
| $e_{es}$                    | The maximum number of executors per slot.                                             |
| Graph.part(n)               | Partition the topology into partitions equal to the number of workers of the topology |
| Cluster.getAvailableNodes() | Get the available nodes in the cluster                                                |
| Nodes.size()                | Total number of nodes in cluster                                                      |
| Partition.size()            | Total number of partitions                                                            |
| Partition. Tasks            | The list of tasks of each partition                                                   |
| MapExecutorToSlots          | Assign tasks to slots                                                                 |

## VI. EXPERIMENTS AND RESULTS

The experimental study is done on the Apache Storm cluster, which has a Nimbus node, Zookeeper node, and two supervisor nodes having three and four slots, respectively. The configuration of the first supervisor is as follows; it has Ubuntu 20.0.1 LTS 64-bit installed on it with GNOME version 3.36.3, with “10.6 GB” memory, Intel@CoreTM i7-4810MQ CPU @ “2.80GHz × 2” processor, and “536.9 GB” disk capacity. The second supervisor has Ubuntu 14.04 LTS with OS type 64-bit with “9.5 GB” memory, Intel@CoreTM i7-4810MQ CPU @ “2.80GHz” processor, and disk capacity “21.7 GB”. Both of the supervisors have 1000 Mb/s network connectivity speed. Each supervisor has Apache Storm 1.1.1, Apache Zookeeper 3.5.7, and Java Open JDK 13. The proposed algorithm of BSS is compared with the default Workload scheduler and the A3 Storm [4] on four benchmark topologies:

1) *SOL topology* [17]: It has a chain-like structure. It has a spout and a set of bolts. It loads its data directly from the data source. Random messages are used to create sentences with words’ lengths specified by the user. It consists of one spout and a user-defined number of bolts. This topology aims to trace the network’s performance, so it is better to keep the computation as minimum as possible.

2) *Rolling count topology* [17]: It applies rolling counts of incoming terms. By the term rolling, it uses a sliding window to trace the statistics of a term until the current window

compared to the one in previous. A rolling count tuple per term is emitted by reaching the end of each time window, and it consists of the term. The term’s rolling count is a metric that points at how this term is trending now and the actual duration of the sliding window. The term can be emitted from more than one node, so a bolt must join and rank the terms. So, it consists of a spout that directly loads the data from the data source, a bolt that splits the sentences, and a rolling count bolt that uses field grouping to count the terms and group them to emit the ranks of each term.

3) *Word count topology* [18]: The spout emits streams of sentences and sends them to a bolt that splits these sentences into words and emits them to another bolt that, using field grouping, can count how many times each word has occurred. Field grouping means that based on the value of the word, the same word must always go to the same instance so that it can be counted.

4) *Spike detection topology* [18]: The spout receives a stream of data from sensors and emits them to bolts to monitor the occurrences of values that have spikes. Spout emits this stream to a bolt named Moving Average, which gets the data grouped according to the IDs of the device. When a new stream of data is received, the bolt aggregates the new values of a device to the list of values of the same device and emits new events consisting of the device ID, its current value, and its values moving average. These emitted events go to another bolt named Spike Detection, as it detects if there is a spike in the current event or not.

## B. Results and Discussion

To evaluate the performance of the proposed BSS algorithm, we consider two performance metrics:

1) *Throughput*: represents the number of tuples processed per unit time [18].

2) *Complete Latency*: is the average time a tuple takes to be entirely processed by the topology [18].

The experiment has been done by applying the proposed BSS, Workload scheduler, and A3 Storm scheduler, the four benchmark topologies mentioned above. Each algorithm has been run three times to get the average results for the three compared algorithms.

3) *SOL topology results*: SOL is generated with one spout and two bolts. The required number of workers is two, and the number of executors and tasks equals a value of nine. The results are depicted in Table III.

As shown in Table III, the Balance scheduling Storm algorithm achieved the highest value of throughput, which was “4059.67 tuples/second” at the second 240. In comparison, the most negligible value of throughput was “15.33 tuples/second” and achieved by the Workload scheduler after “60 seconds”. It is also obvious that the BSS had the best results for the throughput than the two other algorithms till “600 seconds”. At the “660 seconds” and “720 seconds” the A3 Storm showed better throughput results than the BSS and the Workload scheduler algorithms, then starting from the second 780, the

BSS returned to give high throughput than the two other schedulers. Regarding the latency, it is found that the best complete latency was “63.833 milliseconds” for the BSS the first “60 seconds”, while the worst complete latency value was 949, which is recorded by the Workload scheduler after “120 seconds”. The BSS gave the best latency results during the overall execution except at the second 180; the Workload scheduler gave the best latency. Finally, by looking at the average in Table III, it is obvious that the BSS gave better average results in terms of throughput and complete latency.

TABLE III. SOL TOPOLOGY EVALUATION RESULTS

| Time (sec.) | Throughput (tuples/sec.) |                |                  | Complete Latency (Millisecond) |          |                  |
|-------------|--------------------------|----------------|------------------|--------------------------------|----------|------------------|
|             | Workload Scheduler       | A3 Storm       | The Proposed BSS | Workload Scheduler             | A3 Storm | The Proposed BSS |
| 60          | 15.33                    | 84             | <b>183.67</b>    | 437.50                         | 77.50    | <b>63.833</b>    |
| 120         | 867                      | 2045           | <b>1647.67</b>   | 949                            | 764.10   | <b>159.33</b>    |
| 180         | 1364                     | 3384.33        | <b>3194</b>      | <b>443.80</b>                  | 771.80   | 859.47           |
| 240         | 1623                     | 2888           | <b>4059.67</b>   | 387.83                         | 756      | <b>383</b>       |
| 300         | 1136.67                  | 2104.67        | <b>3485</b>      | 401.23                         | 747.67   | <b>251.13</b>    |
| 360         | 538.67                   | 1585.33        | <b>2609.67</b>   | 442.50                         | 794.30   | <b>225.97</b>    |
| 420         | 453                      | 1101           | <b>3333.33</b>   | 515.13                         | 797.30   | <b>197.43</b>    |
| 480         | 212.67                   | 987.33         | <b>1955.33</b>   | 539.63                         | 699.23   | <b>192.63</b>    |
| 540         | 422.33                   | 910            | <b>1871.33</b>   | 549.93                         | 604.87   | <b>186.50</b>    |
| 600         | 637.33                   | 1298.33        | <b>1880.67</b>   | 544.73                         | 489.63   | <b>185.07</b>    |
| 660         | 987.67                   | <b>2072.67</b> | 1865             | 520.07                         | 391.30   | <b>183.37</b>    |
| 720         | 1410                     | <b>1926</b>    | 1804.67          | 484.20                         | 349.80   | <b>184.33</b>    |
| 780         | 2086.33                  | 2044.33        | <b>2781.33</b>   | 445.93                         | 315.27   | <b>181.33</b>    |
| 840         | 1769                     | 2419.33        | <b>3689.33</b>   | 419.10                         | 291.93   | <b>168.97</b>    |
| 900         | 2123.67                  | 2206.67        | <b>2959</b>      | 382.53                         | 276.97   | <b>162.23</b>    |
| Average     | 1043.11                  | 1803.8         | 2487.98          | 497.54                         | 534.2    | 238.973          |

According to the reported results, it is clear that the BSS outperforms both workload scheduler and A3 Storm in both throughput and complete latency according to the overall average values. In contrast, the Workload scheduler and A3 Storm are recorded the worst average on throughput and the complete latency, respectively. Fig. 3 and Fig. 4 depict the comparison between the three algorithms' throughput and complete-time latency results.

4) *Rolling count topology results:* Rolling Count topology is generated on one spout and two bolts with four workers and 26 executors and tasks. The experiments using the three compared schedulers are presented in Table IV and depicted in Fig. 5 and Fig. 6.

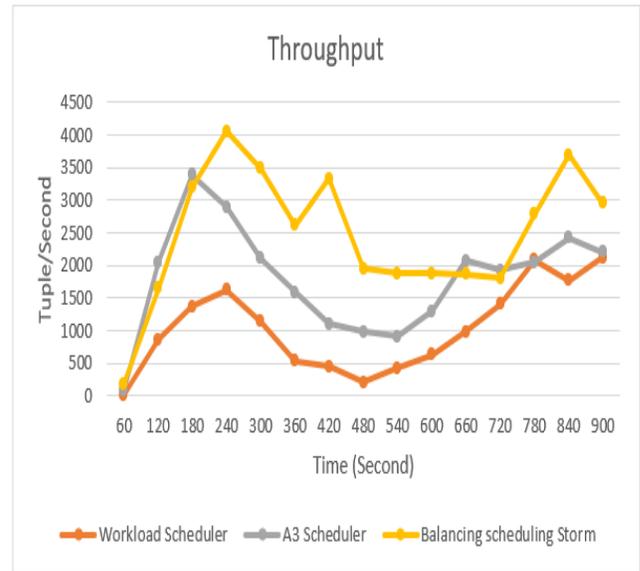


Fig. 3. Throughput Comparison Result between Three Schedulers for SOL.

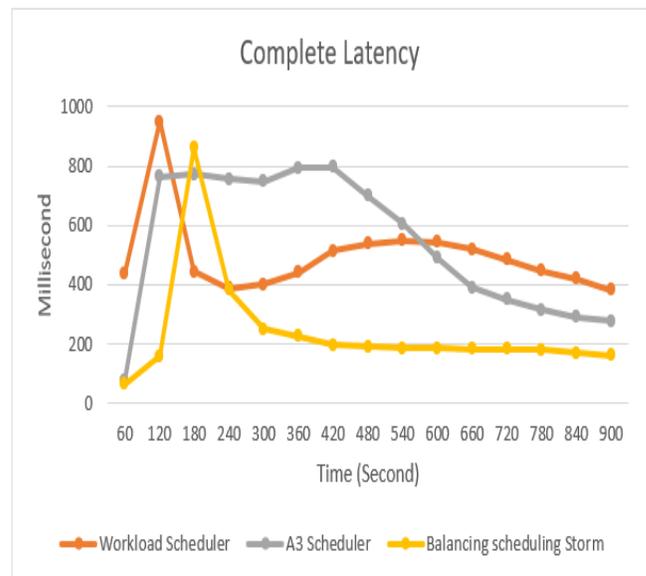


Fig. 4. Complete Latency Comparison Result between Three Schedulers for SOL.

TABLE IV. SOL TOPOLOGY EVALUATION RESULTS

| Time (sec.) | Throughput (tuples/sec.) |          |                  | Complete Latency (Millisecond) |               |                  |
|-------------|--------------------------|----------|------------------|--------------------------------|---------------|------------------|
|             | Workload Scheduler       | A3 Storm | The proposed BSS | Workload scheduler             | A3 Storm      | The proposed BSS |
| 60          | 0                        | 0        | <b>43.67</b>     | 0                              | 0             | 791.38           |
| 120         | 3757.33                  | 2647     | <b>7978.33</b>   | 828.46                         | <b>424.51</b> | 500.12           |
| 180         | 7171.33                  | 4652.33  | <b>11547.33</b>  | 693.91                         | 473.84        | <b>406.36</b>    |
| 240         | 9029                     | 5422.33  | <b>13006</b>     | 650.52                         | 976.38        | <b>366.88</b>    |
| 300         | <b>12457.67</b>          | 5427.67  | 11772            | 501.28                         | 989.39        | <b>374.19</b>    |
| 360         | 11976                    | 4284     | <b>12162</b>     | 452.36                         | 1123.46       | <b>365.63</b>    |
| 420         | 14029                    | 3610.33  | <b>14039.33</b>  | 453.09                         | 1295.57       | <b>360.11</b>    |
| 480         | <b>15489.33</b>          | 2226.67  | 12862.33         | 471.16                         | 1507.76       | <b>348.85</b>    |
| 540         | 13130.33                 | 2218     | <b>17422.33</b>  | 502.71                         | 1740.99       | <b>337.25</b>    |
| 600         | 12256.67                 | 3152     | <b>17847.33</b>  | 562.91                         | 1844.03       | <b>330.63</b>    |
| 660         | 11614.67                 | 5202     | <b>15966.33</b>  | 610.09                         | 1727.88       | <b>331.03</b>    |
| 720         | 10104.67                 | 7001.33  | <b>14791</b>     | 663.46                         | 1605.83       | <b>332.59</b>    |
| 780         | 8298.67                  | 7322.67  | <b>17422.33</b>  | 708.77                         | 1541.14       | <b>328.19</b>    |
| 840         | 11112.67                 | 8103.67  | <b>18629</b>     | 725.97                         | 1527.58       | <b>322.37</b>    |
| 900         | 12185.33                 | 7801.67  | <b>24627.33</b>  | 712.18                         | 1511.18       | <b>310.48</b>    |
| Average     | 10174.18                 | 4604.78  | <b>14007.78</b>  | 569.12                         | 1219.3        | <b>387.07</b>    |

Table IV shows that the BSS has the highest average of throughput, and the A3 Storm has the least average value of throughput. The best value reached of the throughput was “24627.33 tuples/second” at the second 900 by the BSS. During the overall execution the BSS showed better results for throughput. But only at “240 seconds” and “480 seconds”, the

BSS gave lower results and the Workload scheduler showed higher results than the BSS. For the complete latency values, the results showed that the BSS has a minor average of complete latency. In comparison, the Workload scheduler has the highest value of the average complete latency. The least value recorded was “310.48 milliseconds” for the BSS at the second 900, and the highest value was for the A3 Storm “1844.03 milliseconds” at the second 600. Furthermore, as displayed in Table IV, at the first “60 seconds,” both the Workload scheduler and A3 Storm could not finish any data processing, and their recorded throughput was zero. While the BSS processed the data in this limited time and produced “43.67 tuples/second”. At the end of the Table IV, it is shown that the BSS had the best average results for the throughput and complete latency.

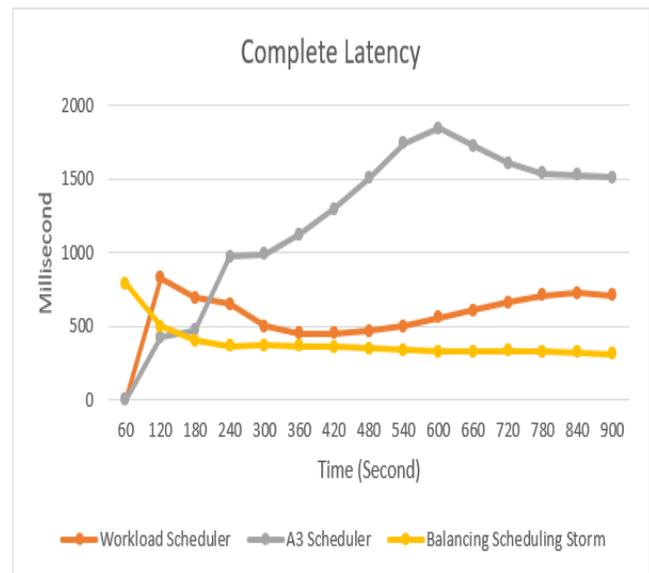


Fig. 5. Complete Latency Comparison Result between Three Schedulers for Rolling Count.

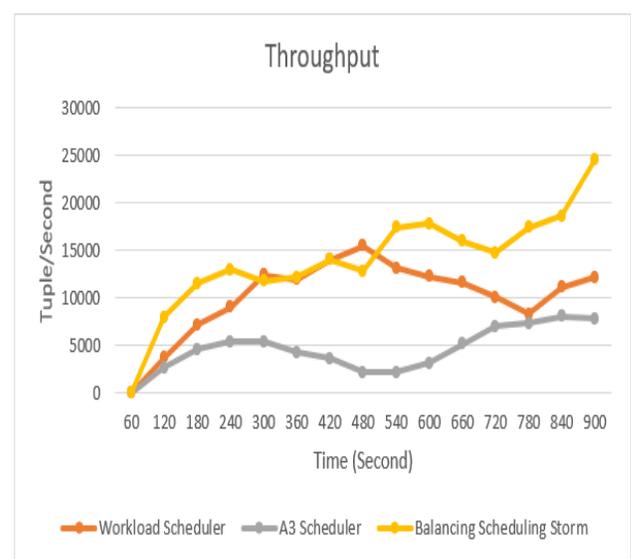


Fig. 6. Throughput Comparison Result between Three Schedulers for Rolling Count.

5) *Word count topology results*: This topology is executed using two workers with one spout and two bolts. It also contains 13 executors and tasks for the “3.8 GB” dataset size.

TABLE V. WORD COUNT TOPOLOGY EVALUATION RESULTS

| Time (sec.) | Throughput (tuples/sec.) |          |                  | Complete Latency (Millisecond) |          |                  |
|-------------|--------------------------|----------|------------------|--------------------------------|----------|------------------|
|             | Workload scheduler       | A3 Storm | The proposed BSS | Workload scheduler             | A3 Storm | The proposed BSS |
| 60          | 417.33                   | 344.67   | <b>2990.33</b>   | 296.67                         | 229.63   | <b>146.87</b>    |
| 120         | 13543                    | 11708.67 | <b>22176</b>     | 96.2                           | 233.87   | <b>84.57</b>     |
| 180         | 27664.33                 | 20740.33 | <b>32432.67</b>  | 60.87                          | 140.53   | <b>52.7</b>      |
| 240         | 26123.33                 | 22706    | <b>29887.33</b>  | 50.97                          | 104.63   | <b>46.93</b>     |
| 300         | 24649                    | 23835    | <b>27501.67</b>  | 46.6                           | 89.83    | <b>44.17</b>     |
| 360         | 22638.33                 | 19374.33 | <b>32028</b>     | <b>44.1</b>                    | 86.03    | 44.73            |
| 420         | 25167                    | 15821.33 | <b>28415.33</b>  | <b>42.43</b>                   | 87.27    | 47.73            |
| 480         | 24097.33                 | 14178.67 | <b>30859.67</b>  | <b>40.63</b>                   | 91.33    | 48.6             |
| 540         | 22573                    | 9098.33  | <b>33317.67</b>  | <b>39.3</b>                    | 97.63    | 50               |
| 600         | 25178.67                 | 8467.33  | <b>29689.33</b>  | <b>37.9</b>                    | 104.77   | 49.93            |
| 660         | 24579                    | 8629.33  | <b>33404</b>     | <b>36.77</b>                   | 109.3    | 49.1             |
| 720         | 12448.33                 | 13077.67 | <b>33584.33</b>  | <b>35.53</b>                   | 110.77   | 48.43            |
| 780         | 19160                    | 16323.67 | <b>27318.33</b>  | <b>35.57</b>                   | 109.1    | 48.47            |
| 840         | 18233                    | 16929.67 | <b>29635.67</b>  | <b>35.53</b>                   | 105.03   | 49.4             |
| 900         | 18468                    | 21994.67 | <b>30709.67</b>  | <b>35.4</b>                    | 99.73    | 50.63            |
| Average     | 20329.31                 | 14881.98 | <b>28263.33</b>  | 62.3                           | 119.96   | <b>57.48</b>     |

Table V shows that the BSS recorded the highest throughput, while the A3 Storm recorded Storm the least. The BSS has reached the maximum value of throughput, which is “33584.33 tuples/second” after “720 seconds,” and the A3 Storm has the least value of throughput, which is “344.67 tuples/second” after “60 seconds”. Regarding the complete latency, the experiments' results showed that the least complete latency value was “35.4 milliseconds” by the Workload scheduler after “900 seconds” of execution. But in contrast, the Workload scheduler recorded the highest value of complete latency, which was “296.67 milliseconds” after the first “60 seconds”. For the BSS, the reported average complete latency value was the least. From the comparison presented in Fig. 7 and Fig. 8, it is obvious that the Balancing Scheduling storm enhanced the system's performance.

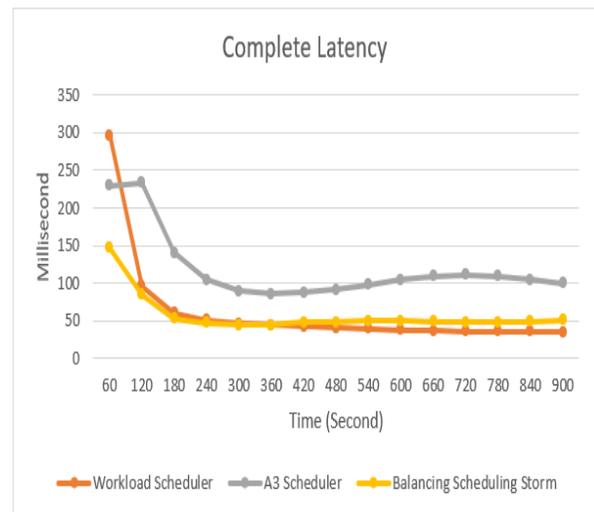


Fig. 7. Complete Latency Comparison Result between Three Schedulers for Word Count Topology.

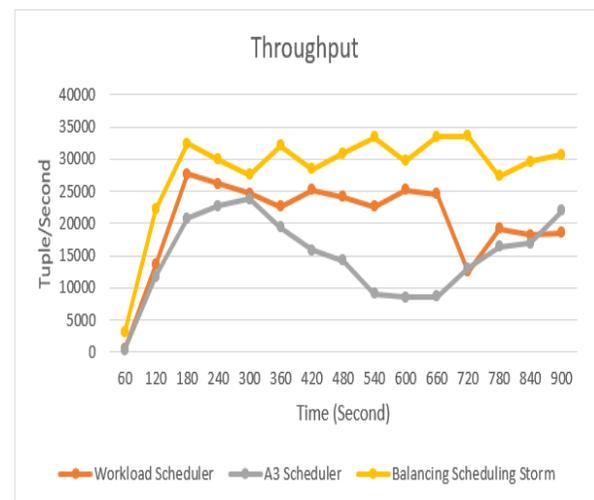


Fig. 8. Throughput Comparison Result between Three Schedulers for Word Count Topology.

6) *Real-time spike detection topology evaluation*: Other experiments are carried out based on the real-time spike detection topology described in the Intel Lab Data [19]. The data set presented here is collected from 54 sensors in the Intel Berkeley Research lab. The data set included about two million readings gathered from these sensors. The topology has been deployed with a dataset about “11 GB” size generated by the repetition of the data presented in Intel Lab Data [19]. The recorded results after running the three schedulers are recorded in Table VI, and the comparison between these results are depicted in Fig. 9 and Fig. 10 are gained.

TABLE VI. SPIKE DETECTION TOPOLOGY EVALUATION RESULTS

| Time (sec.) | Throughput (tuples/sec.) |                |                  | Complete Latency (Millisecond) |               |                  |
|-------------|--------------------------|----------------|------------------|--------------------------------|---------------|------------------|
|             | Workload scheduler       | A3 Storm       | The proposed BSS | Workload scheduler             | A3 Storm      | The proposed BSS |
| 60          | 5.33                     | 0              | <b>55.67</b>     | 1866.24                        | <b>0</b>      | 730.40           |
| 120         | <b>998.67</b>            | 728.33         | 981.33           | <b>797.75</b>                  | 1405.14       | 936.81           |
| 180         | 1383.33                  | <b>1554.33</b> | 1303.33          | 710.82                         | <b>607.35</b> | 697.87           |
| 240         | 1486                     | <b>2358.67</b> | 1955             | 714.35                         | <b>484.22</b> | 513.49           |
| 300         | 1269                     | 1462.67        | <b>2036.67</b>   | 753.16                         | 465.48        | <b>452.48</b>    |
| 360         | 1067                     | 1842.33        | <b>1991.33</b>   | 821.36                         | 476.09        | <b>437.50</b>    |
| 420         | 954                      | 1284.67        | <b>2311.33</b>   | 824.58                         | 475.98        | <b>448.23</b>    |
| 480         | 902.67                   | 903.33         | <b>1740.33</b>   | 813.78                         | 531.24        | <b>481.70</b>    |
| 540         | 671.33                   | 1143.67        | <b>1270.67</b>   | 789.39                         | 524.58        | <b>515.08</b>    |
| 600         | 1006                     | 999            | <b>1578</b>      | 786.27                         | 535.59        | <b>488.90</b>    |
| 660         | 745.67                   | 1357.67        | <b>1746.67</b>   | 778.69                         | 525.92        | <b>473.79</b>    |
| 720         | 920                      | 1397.33        | <b>2205.67</b>   | 776.01                         | 515.26        | <b>462.69</b>    |
| 780         | 1141.67                  | 1643.33        | <b>1747.33</b>   | 784.31                         | 504.56        | <b>461.25</b>    |
| 840         | 1486.67                  | 1553.33        | <b>1889.33</b>   | 786.86                         | 503.45        | <b>444.81</b>    |
| 900         | 2056.67                  | 1630           | <b>2269.33</b>   | 709.07                         | 501.22        | <b>427.87</b>    |
| Average     | 1072.93                  | 1323.91        | <b>1672.13</b>   | 847.51                         | 537.07        | <b>531.53</b>    |

Table VI shows the comparison results between the three algorithms in terms of throughput and complete latency. The Balancing Scheduling Storm has the maximum value of the average throughput. In contrast, the Workload scheduler has the minimum value. The best result of the throughput is “2311.33 tuples/second”, which was reached by the Balancing scheduling storm after “420 seconds”. The smallest value was “0 tuples/second,” which was reached by the A3 Storm after “60 seconds”. The next worst value of throughput was “5.33 tuples/second” at the first “60 seconds” and was reached by the Workload scheduler.

Also, the latency results are described in Table VI. The worst latency value was for the Workload scheduler at the first minute; it was “1866.24 millisecond”. The best value was for the Balancing Scheduling storm after “15 minutes” it reached the value of “427.87 milliseconds”. And by looking at the end of the Table VI, it is obvious that the BSS had the best average results in both terms, the throughput and the complete latency.

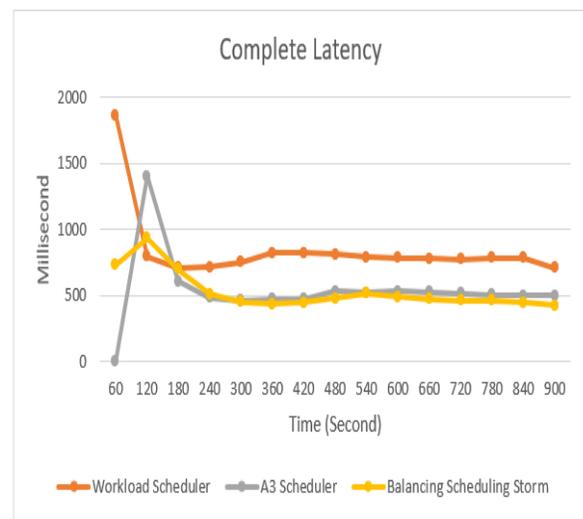


Fig. 9. Complete Latency Comparison Result between Three Schedulers for Spike Detection Topology.



Fig. 10. Throughput Comparison Result between Three Schedulers for Spike Detection Topology.

## VII. CONCLUSION AND FUTURE WORK

In this paper, the Balance Scheduling on Storm (BSS) scheduler is proposed. It is a hybrid scheduling algorithm based on two existing scheduling algorithms: the Workload scheduler and the A3 Storm. It takes balancing the workload between the cluster nodes while minimizing the communication network between them and the inter-executor traffic. This proposed algorithm has been compared with the two existing schedulers using two metrics, throughput, and complete latency metrics. The BSS algorithm has shown better performance. The results show high throughput more than the other two algorithms and less latency. This has improved the performance of the system.

The comparison done in this paper used two essential metrics: throughput and complete latency. In future work, an enhancement in the performance will be carried out for other metrics, such as the amount of memory and resources used to give better performance and enhancement than the already given performance in this thesis. The comparison can be done with more scheduling algorithms than the two algorithms already compared with. New research could enhance the number of resources and the amount of memory and their usage instead of the balanced scheduling. The experiments would take place on more topologies to check their suitability and performance.

## ACKNOWLEDGMENT

The authors acknowledge the Academy of Scientific Research and Technology in Egypt for its support and funding this paper within the Academy project no.: 6490.

## REFERENCES

- [1] Y. Riahi and S. Riahi, "Big Data and Big Data Analytics: Concepts, Types and Technologies," *International Journal of Research and Engineering*, vol. 9, no. 5, pp. 524-528, 2018.
- [2] A. C. Lyons and J. Grable, "An Introduction to Big Data," *JOURNAL OF FINANCIAL SERVICE PROFESSIONALS*, vol. 72, no. 5, pp. 17-20, 2018.
- [3] L. Aniello, R. Baldoni and L. Querzoni, "Adaptive Online Scheduling in Storm," In DEBS 2013, 2013.
- [4] A. Muhammad and M. Aleem, "A3 Storm: topology , traffic , and resource aware storm," *The Journal of Supercomputing*, no. 77, p. 1059–1093, 2021.
- [5] R. Mall, *Introduction. Real-Time Systems (Theory and Practice)*, Kharagpur : Dorling kindersley (India), 2007.
- [6] K. J. Giri and T. A. Lone, "Big Data - Overview and Challenges.," *International Journal of Advanced Research in Computer Science and Software Engineering*, pp. 525-528, 2014.
- [7] G. Karypis and V. Kumar, "Multilevel k-way Partitioning Scheme for Irregular Graphs," *JOURNAL OF PARALLEL AND DISTRIBUTED COMPUTING*, vol. 48, no. 1, p. 96–129, 1998.
- [8] L. Fischer and A. Bernstein, "Workload Scheduling in Distributed Stream Processors using Graph Partitioning," in *2015 IEEE International Conference on Big Data (IEEE BigData 2015)*, Santa Clara, CA, USA, 2015.
- [9] A. Jain, *Mastering Apache Storm*, 1st ed., Packt Publishing, 2017.
- [10] S. Saxena and S. Gupta, *Practical Real-Time Data Processing and Analytics*, Packt Publishing, 2017.
- [11] B. Peng, M. Hosseini, Z. Hong, R. Farivar and R. Campbell, "R-Storm: Resource-Aware Scheduling in Storm," in *ACM Middleware '15 Proceedings of the 16th ACM Annual Middleware Conference*, Vancouver, Canada, 2015.
- [12] J. Xu, Z. Chen, J. Tang and S. Su, "T-Storm: Traffic-aware Online Scheduling in Storm," in *2014 IEEE 34th International Conference on Distributed Computing Systems*, 2014.
- [13] C. Li, J. Zhang and Y. Luo, "Real-time scheduling based on optimized topology and communication traffic in distributed real-time computation platform of storm," *Journal of Network and Computer Applications*, vol. 87, pp. 100-115, 2017.
- [14] X. Liu and R. Buyya, "D-Storm: Dynamic Resource-Efficient Scheduling of Stream Processing Applications," in *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*, Shenzhen, China, 2017.
- [15] A. Muhammad, M. Aleem and M. A. Islam, "TOP-Storm: A topology-based resource-aware scheduler for Stream," *Cluster Computing*, no. 24, pp. 417-431, 2021.
- [16] G. Karypis and V. Kumar, "METIS—A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes and Computing Fill-Reducing Ordering of Sparse Matrices," *University of Minnesota, Department of Computer Science and Engineering, Army HPC Research Center*, 1998.
- [17] B. Gautam and A. Basava, "Performance prediction of data streams on high-performance architecture," *Human-centric Computing and Information Sciences*, vol. 9, no. 2, 2019.
- [18] M. V. Bordin, *A benchmark suite for distributed stream processing systems*, 2017.
- [19] S. Madden, "Intel Lab Data," 2004. [Online]. Available: <http://db.csail.mit.edu/labdata/labdata.html>. [Accessed 5 July 2021].

# Extract Concept using Subtitles in MOOC

Aarika Kawtar\*, Habib Benlahmar, Mohamed Amine Naji, Elfilali Sanaa, Zouheir Banou  
Laboratory of Information Technology and Modeling, Hassan II University of Casablanca, Faculty of Sciences  
Casablanca, Morocco

**Abstract**—Massive open online courses (MOOCs) are a variety of courses offered through the online mode, paid or unpaid and has evolved as an excellent learning resource for students. The structure of the course design is mainly linear where there are a few video lectures provided by either professors of several universities, or people with expertise in the particular subject. They are usually graded on a weekly basis through quizzes or peer-graded assignments. The objective of this paper is to extract the concepts taught in the videos from the subtitles, which could later be used to enhance recommendations of the learners using their clickstream data. The teachers could also use this to see the demand for their courses. Evaluate two keyword extraction methods, which are BERT and LDA using different Coursera courses. The experimental results show that BERT outperforms LDA in terms of Coherence.

**Keywords**—LDA; BERT; topic coherence; overlap coefficient

## I. INTRODUCTION

One of the most profitable and in demand businesses in today's world are those of Massive open online courses (MOOC). Not only do they offer a vast range of lectures on almost all the topics be it the medical field, or some complex lessons on coding, but can also be easily accessible by everyone sitting at home [1]. These platforms have attracted a large number of people, which sums up to nearly 10 million participants from all over the world. Coursera, Udemy and EdX are some of the classic examples of MOOC [2]. The format of these platforms is similar, where professors or trained people share video lectures covering a particular topic. They use different methods of teaching like using powerpoint presentations, whiteboard or even the electronic boards. Some platforms invest a lot in making their videos interesting and visually appealing, hence they incorporate graphics and colorful animations. This helps in drawing more attention from the learners, especially the younger crowds. The lectures are usually grouped into few modules and a set of modules makes up a course. The module system helps the learner keep track of their progress and also have a better understanding of the pre requisites. The teachers find the module systems easier as it is easier to set assignments and other assessment related work. Each module usually runs for a week, however, it totally depends on the viewers' interests. Weekly deadlines are set, which are flexible. This means that an ideal schedule is provided, which if followed thoroughly can benefit an average learner. However, it is their choice ultimately on how much time they want to spend on it, it could be earlier than the target date of even later. At the end of each module, there is an assessment held. There are several ways in which one is assessed to see how much of the course they have grasped. Some of these assessment techniques are quizzes, projects and peer-graded assignments. Few courses even have cutoffs to be

cleared at the end of module assessment. Failure to complete this successfully would not permit the learner to proceed to the next module or it might not consider the module as complete [3]. Upon course completion, the learner receives a certificate of completion from the institute offering it and it can be considered as a legitimate proof of knowledge acquired, and can be updated in resumes and professional profiles. Courses that have strict assessments do not provide the certificates until all the quizzes have been cleared with the minimum required cutoff and all the peer graded assignments have been checked by the required number of co learners. The legitimacy of MOOC has gone so far that nowadays, universities offer these courses as electives as proper curriculum courses with college credits awarded on their completion. The college provides these courses and has their own assessment methods, however, the students have to complete these courses through the platform in order to receive the assigned number of credits. These courses can be free, but mostly they have to be purchased. Another alternative provided by MOOC is that some courses can be audited for free but do not provide completion certificates hence the purpose is solely for acquiring knowledge.

MOOC provides a form of social learning where interactions constantly take place between learners and the teachers. It paves way for mass learning and personalised comprehension. Even though there is no face-to-face communication taking place, these platforms have been successfully been able to break the barriers of any type of communication hindered otherwise [4]. There are different ways one can engage themselves with the platforms. Learners and teachers can both participate in forum and discussions, helping fellow learners and students. Some even start taking lectures of their own. Others work on in video editing, as mentioned earlier, adding good graphical depictions of what is being explained or colourful animations. There is a lot work that has to be dealt in the back end of the sites or apps belonging to these platforms. A large group of people also contributes by providing constructive feedback and suggests improvements. These are constantly monitored and taken note of in order to improve the user interface of their platforms and attract more learners to purchase their products. These learning methods are completely different from the physical mode of learning and open a wide door of new opportunities to explore [5]. Hence, we can say that the most important factor which determines the success of these MOOCs is the engagement of the students, however not a lot of research has been carried out on how the student engagement affects the platforms. All MOOC platforms primarily run on how much they have been used and a decline in student engagement can give a massive blow to these businesses. It is of utmost importance that the

\*Corresponding Author

engagement is always constantly monitored and changes being continuously implemented in order to keep them high [6]. The discussion forums play a vital role in checking engagement, along with website visits, registrations, clicks etc. However, it is not an easy task to keep track of the engagement as there are so many parameters that have to be taken into consideration while doing the analysis. Some of them are course enrollments, course completion, discussion forums, etc. [7].

These courses come with their own set of disadvantages. Though they attract a large number of student registrations, recent studies have shown that only a small fraction of these students complete their courses [8]. According to statistics provided by Coursera, almost 75% of the courses enrolled by students have not been completed [9]. Another problem is that these platforms do not come with keyphrases and it is going to be a laborious task to identify them manually and will take up a lot of time. This means that one cannot search for courses based on particular topics. There are a variety of topics mentioned in each video, but there is no way of keeping track of these. It is important to do so as it can help recommending better courses to those who show interest in topics. Topic based searches can be made than course-based searches and it will be easier for the learner to choose their apt course based on how much do the topics cover in the course line up with their topics of interest.

Keyphrases are important and significant expressions consisting a collection of words. They give us the contents of the data, or even sometimes summarize it [10]. There have been several algorithms developed to extract keywords from scripts, notes etc. These are used in data mining like clustering of documents, providing recommendations and formulation of queries [11], [12]. Bidirectional Encoder Representations from Transformers (BERT) is one of the models that can be used for keyphrase extraction. This model is used to make sequential recommendations based on past data. The distinctive feature of this method is that it can incorporate context from both sides, unlike other sequential predictors, which only do it from left to right [13]. Latent Dirichlet allocation (LDA) model is a probabilistic modeling algorithm. It is commonly used to identify the topics in a collection of texts. It is usually used in image retrieval and face recognition technologies [14], [15].

Instructors face problems in analyzing each student's level of understanding in order to improve the quality of courses or to provide referral systems. Although the number of students enrolling in courses has increased, very few of them actually complete the course. Therefore, it is necessary to track learner journey data to know what interests them. The goal of this paper is to extract the concepts taught in the videos from the subtitles, which could then be used to improve the learners' recommendations using their path data. Instructors could also use this to learn about the demand for their courses.

In this paper, we have attempted to extract concepts from the subtitles of video lectures of courses offered by Coursera using BERT and LDA models for key phrase extraction. A comparison is made between the results obtained by both.

The paper is organized as follows. In Section 2, we review related work on concept extraction.

Section 3 is devoted to the context of our experimental study, detailing the dataset collected from the Coursera MOOC videos and the models (LDA and BERT) that we will use for this study.

In Section 4, we show our proposed algorithms for concept extraction from the sub-titling of the experimental results which show a better concept extraction. In section 5 we end with a conclusion that shows the results of our work.

## II. RELATED WORK

In our study, we have tried to automatically extract keyphrases from the subtitles of the videos. In general, there are two ways in which these extractions are carried out [16]. The first approach is supervised, where there is binary segregation of each word into either keyphrase or not a keyphrase [17]. The second approach is unsupervised. In this approach, the words are ranked based on what the algorithm asks it to do, for example probability of occurrence, or even usage in the course. Some commonly used machine learning algorithms are Naïve Bayes and support vector machines [18].

Yi-fang et al developed an algorithm called KIP algorithm. In this algorithm the extracted words were first examined and scored on the basis of three factors. The first factor was their frequency in the text. This means they checked how frequently the word occurred in the text. Second parameter considered was their specificity. This means there is a check on how specific or unique the words are to the course provided. This information is also gathered by checking on the neighborhood data. Last parameter taken into consideration is its contents, as in the words that are related to the examined word. The words are arranged in order of their scores. The words that obtain high scores are later categorized as keyphrases [12]. Another similar type of work can be seen in Xiaojun et al's paper used information from the neighborhood documents to get more data and then this data was graphically represented along with the data of the document where keywords need to be extracted. These data were compared and the keyphrase were extracted accordingly [10]. A very similar study to ours was found in the works of Raga et al. They used this method to navigate to the exact part of the video, or access a video segment by just searching for the keyword. They considered factors like statistical and visual features while implementing the algorithm [19]. A model called Text Rank was developed by Rada and Paul where they took a graphical approach to rank the words [20]. The TPR (Topical Page Rank) approach is another one proposed by Liu et al where first the segregation occurs based on various topics and then the TPR algorithm is individually run on each one [18]. Some other algorithms were developed based on using deep learning [21], frequency of occurrence of words [22], word embedding vectors and graphical ranking [23].

Our study draws inspiration from all these works, but still manages to stand apart as we aim on increasing engagement by giving personalised recommendations to learners based on their search history or clickstream data. To ensure this, keyphrases have been extracted from the subtitles using two algorithms and the best of these two on experiments could be used to develop better recommendation systems. Elaborating on that, these keyphrases can be used to match with the learner's

interests thus giving better course recommendations. The teachers can also benefit from this study. The clickstream data allows teachers to know which topics are more in demand and will encourage them to record lectures covering those topics. This will help the algorithm detect their courses and recommend it to the learners. They can also gauge the learner engagement and see which part of their courses attract more attention.

### III. METHODOLOGY

#### A. Datasets

The study will use the dataset called "MOOC DATA". This dataset has been derived from the subtitles of the course videos from Coursera platform. All the words are split up into individual components and these could further be sent into algorithms for keyphrase extraction. This dataset consists of a total of four folders named:

- "CSEN" – Computer Science in English
- "CSZH"- Computer Science in Chinese
- "EcoEN"- Economy in English
- "EcoZH"- Economy in Chinese

The statistics of the four datasets are listed in Table I, where #courses, #videos, are the total number of courses, videos, in each dataset.

TABLE I. DATASET STATISTICS

| Dataset | Domain           | Language | #courses | #video |
|---------|------------------|----------|----------|--------|
| CSEN    | Computer Science | English  | 18       | 2,849  |
| EcoEN   | Economics        | English  | 5        | 381    |
| CSZH    | Computer Science | Chinese  | 8        | 690    |
| EcoZH   | Economics        | Chinese  | 8        | 455    |

However, for the sake of better understanding and better research, only the "CSEN" folder was used for the study. This folder contained two JSON files, one of these files was called "candidates" and the other one was called "captions". Again, since the aim of our study only deals with the subtitles of the videos, only the caption file was utilized. This table contained the video captions of 18 computer courses; the size of this file is 216 MB. Table II shows how the subtitles were sliced and stored.

The first column is usually ignored as it is the serial numbers. The second column is the Course ID. This is a unique code, which is used to identify the particular course. For the course we considered (Computer Science with English), the course id is 1. The column next to this is the text. It consists of the script of the subtitles and is called transcript. This script is so precise that it also has details like parts of the video where there is music. That's what makes the process of keyword extraction challenging. The music is used way too many times; the system might mistake it to be a keyphrase while we know it is just the background music being referred. Hence it is important that all these unwanted parameters are taken care of

at the initial stages and the algorithm is not affected due to them. The column next to it is the tagged column. Here, we see that every word has been sliced up (including the music). The 3 gram model is used to carry out this process. For example, the first row shows that the text has been separated into tags like.

"MUSIC", "Today", "we", "re" and so on. Again, on running models, the music words should not be considered for keyphrase extraction. The last column is video id which is the unique number given to the video in a course, and is used to refer to a particular course. The course is the same while the videos from which the words were extracted from were different. Our study uses the first 5 videos from the course with course id 1. This dataset will run through two models and a comparison will be drawn regarding which is the better one to consider for keyphrase extraction.

#### B. Pre-Processing

Pre-processing is the process where the raw data received is converted into a form that is comprehensible and useful. It is extremely crucial to ensure that data pre-processing has been done before carrying out any analytical task [24] [25]. This helps in having a dataset of good quality. Process used to split the text or segmenting a text to words, meaningful parts or phrases is called tokenization. In this process, punctuations, whitespaces and other non- alphanumeric characters are not considered, all characters are converted to lowercase and stopwords (conjunctions, articles, etc.) are removed [26].

Before proceeding, there is another concept that needs to be looked upon, which is an n-gram. An n-gram (or Q gram) is basically a sliced part of a longer string consisting of n characters. They are usually obtained from a sample text or some form of speech [27]. It could be words, phrases, letters sometimes even syllables. It is a very efficient means of implementation. On conversion in n-grams the string gets embedded into a vector and is further compared with other data of similar type. Its consistency and distribution can be measured too [28]. An n-gram model is a probabilistic language model, where it is used to make predictions of the items succeeding it in the form of a sequence known as an (n-1)- order Markov model. These models find their extensive usage in computational linguistics, communication theory and data compression. There are two major advantages of using thesen-gram models and algorithms. One of them is simplicity, the model is comparatively simpler to operate and execute than its other counterparts. Secondly, its scalability is a boon. At higher n values, this model is able to store more contexts with a space- time tradeoff which has been understood well. This allows the smaller experiments to efficiently expand.

In our study data has been obtained beforehand from the dataset in order to run it with various algorithms. The words from the video subtitles have been sliced out into different words. Each of these go through the algorithm to obtain results on whether it is a keyword or not. That is decided based on other data like how frequently the word is used or its significance in the text. This process will help identify the key topics covered in the course. The data was retained from the dataset, however, unnecessary information like the tagged column and stopwords were eliminated altogether and n-grams were generated.

### C. Models used

1) **BERT**: Bidirectional Encoder Representations from Transformers (commonly known as BERT) is a machine learning model that is used for language representation [28]. These models are pre-trained and they force the model to study the semantic data in between and withing the sentences. Unlike other similar models which only function from left to right, BERT works from both directions i.e. it is bidirectional, just as its name says [29]. This algorithm takes the final hidden state of the first token and uses it to represent the entire sequence for tasks which require classification of texts. When BERT is incorporating with another output layer, there is an advantage of minimal number of parameters being necessary to be learnt from scratch [30]. There is a particular format that any input data needs to fulfill if it has to undergo the BERT model. A special token which consists of the special classification embedding called [CLS] is put prior to every sentence to fulfill this criterion. Another special token that is used is the [SEP]. It is placed at the end of each and every sentence in order to make a clear separation between the segments [26]. BERT also relieves the problem of masked language model (MLM), where it randomly covers some of the input and expects the algorithm to predict it based on the date of the surrounding words. The next sentence prediction (NSP) is also used. Fine-tuning techniques are of various kinds based on how much of the architecture needs to be trained [31]. Basically, it is a sequential predictor. Google uses BERT to enhance its search predictions. In our first study, we have taken the data from the dataset and run it with BERT model [35].

For BERT analysis the probability analysis can be represented using the following language model by Equation (1) [36]:

$$P(w_1, w_2, \dots, w_T) = \prod_{t=1}^T P(w_t | w_1, w_2, \dots, w_{t-1}) \quad [36] \quad (1)$$

Where  $w_1, w_2, \dots$  are the different individual entities of which we need to find the probability distribution and  $T$  is the total number of entities. In our case, this is the probability distribution of each word in the subtitle file.

2) **LDA**: The Latent Dirichlet Allocation (or LDA) is a probabilistic model. The main aim of this model is to represent documents as different topics and each of these topics are characterised by a distribution over words [32]. The assumption made here is that every course has a set of topics already and the text (subtitles in our case) have relevant information to summarise these topics and hence, they can be grouped under them. The algorithm tells us the similarities in the data by grouping them into common topics [14]. It gives us a distribution of the word usage and when we search for a particular word, it refers to this distribution [33]. Supervised Machine Learning algorithms are used to run the model. This approach is used as a solution to a lot of problems related to topic identification, face recognition, web spam classification

and entity resolution [34]. The second part of our study deals with LDA.

To find the normal probability density function using the LDA method, the formula is given by Equation (2) [37]:

$$P(X|\pi_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left[ -\frac{1}{2} (X - \mu_i)' \Sigma^{-1} (X - \mu_i) \right] \quad [37] \quad (2)$$

Where,

$\pi_i$  – Probability density function

$x$  – Multivariable normal

$\mu_i$  – Mean vector

$\Sigma$  – Variance-covariance matrix.

This can be used if all the matrices for all the populations are homogenous. The decision rule of the LDA algorithm is based on the Linear score function, which is defined by Equation (3):

$$S_i^L(x) = -\frac{1}{2} \mu_i' \Sigma^{-1} \mu_i + \mu_i' \Sigma^{-1} X + \log P(\pi_i) \quad [38] \quad (3)$$

Where following substitutions are made:

$$d_{i0} = -\frac{1}{2} \mu_i' \Sigma^{-1} \mu_i \quad ; \quad d_{ij} = j\text{th element of } \mu_i' \Sigma^{-1}$$

$d_i^L(X)$  is the linear discriminant function (4) i.e.

$$d_i^L(X) = d_{i0} + \sum_{j=1}^p d_{ij} x_j \quad (4)$$

Therefore we get Equation (5) [38],

$$S_i^L(x) = d_i^L(X) + \log P(\pi_i) \quad [38] \quad (5)$$

## IV. EXPERIMENT AND RESULTS

### A. BERT

BERT analysis was first carried out on the preprocessed data. There was a restriction put on the number of concepts that could be extracted to only 3 concepts per line. The n-gram set for each concept was between 1 and 3. Fig. 1 depicts the results obtained.

Fig. 2 shows the coherence and the average overlap of the topics when the data was processed through the BERT model. 20 topics were given to derive BERT's selected keywords. The topic coherence graph shows linear increase upto topics, which is also followed by a linear increase, but the slope gets reduced. The average topic overlap graph shows a steep linear decrease initially up to 2 topics, after which the slope reduces. Finally after 3 topics, the line almost flattens out. Both the graphs overlap at a point in the earlier stages. The ideal number of topics is 4.

```
Word : standard_template_library default_constructor memory_address - Score : 0.5418

Word : random_number_generation standard_template_library standard_template_library - Score : 0.641

Word : type_safe standard_library type_safety - Score : 0.6423

Word : systems_implementation_language object_oriented native_types - Score : 0.7735

Word : standard_template_library standard_template_library standard_template_library - Score : 0.6942

```

Fig. 1. The Concept Extracted from Subtitles with BERT.

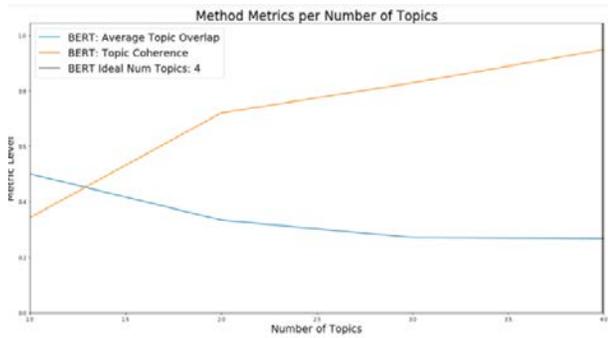


Fig. 2. BERT Coherence Score with Overlap Coefficient.

### B. LDA

The second studies were carried out using the LDA model. Again, concepts per line were restricted to 3 and the n-gram was set between 1 and 3. Fig. 3 gives us the results obtained.

```
singl macro topic
prefer initi transit
build dealloc data_structur

morph discrimin reinvent
basic deal topic
answer rest processor

function display store
resourc bubblesort discrimin
stay pure core

mean basic involv
bell head safe_cast
support referenti save

short argument chang
paradigm treat xerox
assign opportun obsolet

```

Fig. 3. The Concept Extracted from Subtitles with BERT.

Fig. 4 gives us an insight of the results for the same. A graph containing coherence, average overlap of topics was plotted where 20 topics were given to derive the LDA selected keywords. The graph of Topic Coherence shows a peculiar trend. It remains constant throughout and shows no variations at all. The average topic overlap shows a slight decreasing linear trend up to 2 topics, then it remains constant and above 3 topics there is a further linear slight decrease. The overall overlap decrease is very small. Unlike what we observed in the graph of BERT analysis, in this graph we do not see any intersection of the two parameters.

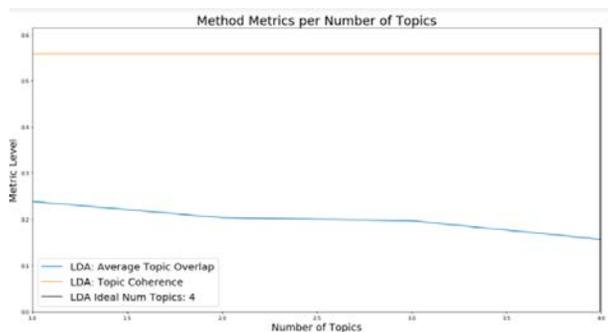


Fig. 4. LDA Coherence Score with Overlap Coefficient.

### C. BERT v/s LDA

In order to draw comparisons with both the studies, the following parameters were considered.

1) *Overlap coefficient*: It is the measure of similarity that is used to track the amount of overlap between two finite sets. In other words, we can say that it is the intersection of two sets. Our studies showed the average overlaps of the topics in LDA to be higher than that of BERT analysis.

2) *Topic coherence*: It measures the total score of a single topic by measuring the degree of semantic similarity between the high scoring words of the topic. The consistency of the concepts by BERT was found to be higher than that of LDA.

As consistency is the more prioritised factor, overall, it can be concluded that BERT is the better model to use for keyphrase extraction of video subtitles in MOOC than LDA as it gives us clearer information about the topic coherence.

### V. CONCLUSION AND FUTURE PROSPECTS

Our studies show that BERT was a better model that could be implemented in order to extract keyphrases from the video subtitles from MOOC videos. The MOOC industry is booming and will continue to do so in the future. It is important to ensure that the course completion rate is high. Now that one can identify the key topics in a course using BERT model, further programming can be done to link these results with the search history of the learner. When any of the key topics are searched, these courses should show up and similar courses be recommended. This will ensure that the learner finds exactly what they are looking for thus motivating them to complete the course and enjoy it. This also helps give them personalised recommendations. As mentioned earlier, the teachers recording the courses also will vastly benefit from this. They can check the engagement of the students in their courses, or have an idea about which part of their video is watched more or gets more demand. They can also use this data to record lectures accordingly so that their courses appear on the top of the recommendations.

### REFERENCES

- [1] Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom research into edX's first MOOC. *Research & Practice in Assessment*, 8, 13-25.
- [2] Zhou, Y., & Xu, Z. (2020, August). Multi-Model Stacking Ensemble Learning for Dropout Prediction in MOOCs. In *Journal of Physics: Conference Series* (Vol. 1607, No. 1, p. 012004). IOP Publishing.
- [3] Brinton, C. G., & Chiang, M. (2015, April). MOOC performance prediction via clickstream data and social learning networks. In *2015 IEEE conference on computer communications (INFOCOM)* (pp. 2299-2307). IEEE.
- [4] Brinton, C. G., Chiang, M., Jain, S., Lam, H., Liu, Z., & Wong, F. M. F. (2014). Learning about social learning in MOOCs: From statistical analysis to generative model. *IEEE transactions on Learning Technologies*, 7(4), 346-359.
- [5] Guo, P. J., Kim, J., & Rubin, R. (2014, March). How video production affects student engagement: An empirical study of MOOC videos. In *Proceedings of the first ACM conference on Learning@ scale conference* (pp. 41-50).
- [6] Kuh, G. D. (2003). What we're learning about student engagement from NSSE: Benchmarks for effective educational practices. *Change: The magazine of higher learning*, 35(2), 24-32.

- [7] Giannakos, M. N., Sampson, D. G., & Kidziński, Ł. (2016). Introduction to smart learning analytics: foundations and developments in video-based learning. *Smart Learning Environments*, 3(1), 1-9.
- [8] Mohamad, N., Ahmad, N. B., & Sulaiman, S. (2017). Data preprocessing: a case study in predicting student's retention in MOOC. *Journal of Fundamental and Applied Sciences*, 9(4S), 598-613.
- [9] Bach, S. H., Broecheler, M., Kok, S., & Getoor, L. (2010). Decision-driven models with probabilistic soft logic.
- [10] [Wan, X., & Xiao, J. (2008, July). Single Document Keyphrase Extraction Using Neighborhood Knowledge. In *AAAI* (Vol. 8, pp. 855-860).
- [11] Gollapalli, S. D., & Caragea, C. (2014, June). Extracting keyphrases from research papers using citation networks. In *Twenty-eighth AAAI conference on artificial intelligence*.
- [12] Wu, Y. F. B., Li, Q., Bot, R. S., & Chen, X. (2005, October). Domain-specific keyphrase extraction. In *Proceedings of the 14th ACM international conference on Information and knowledge management* (pp. 283-284).
- [13] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019, November). BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 1441-1450).
- [14] Krestel, R., Fankhauser, P., & Nejdl, W. (2009, October). Latent dirichlet allocation for tag recommendation. In *Proceedings of the third ACM conference on Recommender systems* (pp. 61-68).
- [15] Canini, K., Shi, L., & Griffiths, T. (2009, April). Online inference of topics with latent Dirichlet allocation. In *Artificial Intelligence and Statistics* (pp. 65-72). PMLR.
- [16] Hasan, K. S., & Ng, V. (2014, June). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1262-1273).
- [17] Hasan, K. S., & Ng, V. (2010, August). Conundrums in unsupervised keyphrase extraction: making sense of the state-of-the-art. In *Coling 2010: Posters* (pp. 365-373).
- [18] Albahr, A., Che, D., & Albahar, M. (2021). A novel cluster-based approach for keyphrase extraction from MOOC video lectures. *Knowledge and Information Systems*, 63(7), 1663-1686.
- [19] Koka, R. S., Chowdhury, F. N., Rahman, M. R., Solorio, T., & Subhlok, J. (2020, December). Automatic identification of keywords in lecture video segments. In *2020 IEEE International Symposium on Multimedia (ISM)* (pp. 162-165). IEEE.
- [20] Mihalcea, R., & Tarau, P. (2004, July). TextRank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).
- [21] Martinez - Romo, J., Araujo, L., & Duque Fernandez, A. (2016). S em G raph: Extracting keyphrases following a novel semantic graph - based approach. *Journal of the Association for Information Science and Technology*, 67(1), 71-82.
- [22] Florescu, C., & Caragea, C. (2017, July). Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1105-1115).
- [23] Pan, L., Wang, X., Li, C., Li, J., & Tang, J. (2017, November). Course concept extraction in moocs via embedding-based graph propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 875-884).
- [24] Ramírez-Gallego, S., Krawczyk, B., García, S., Woźniak, M., & Herrera, F. (2017). A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, 239, 39-57.
- [25] Terrizzano, I. G., Schwarz, P. M., Roth, M., & Colino, J. E. (2015, January). Data Wrangling: The Challenging Journey from the Wild to the Lake. In *CIDR*.
- [26] Mifrah, S., & Benlahmar, E. H. (2020). Topic modeling coherence: A comparative study between LDA and NMF models using COVID'19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, 5756-5761.
- [27] Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175).
- [28] Li, W. J., Wang, K., Stolfo, S. J., & Herzog, B. (2005, June). Fileprints: Identifying file types by n-gram analysis. In *Proceedings from the Sixth Annual IEEE SMC Information Assurance Workshop* (pp. 64-71). IEEE.
- [29] Jawahar, G., Sagot, B., & Seddah, D. (2019, July). What does BERT learn about the structure of language?. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.
- [30] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8, 842-866.
- [31] Khodeir, N. A. (2021). Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT. *IEEE Access*, 9, 58243-58255.
- [32] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [33] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3, 993-1022.
- [34] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1), 5228-5235.
- [35] Bíró, I., Siklósi, D., Szabó, J., & Benczúr, A. A. (2009, April). Linked latent dirichlet allocation in web spam filtering. In *Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web* (pp. 37-40).
- [36] Dadure, P., Pakray, P., & Bandyopadhyay, S. (2021). BERT-Based Embedding Model for Formula Retrieval. *CLEF*.
- [37] Jin, Q., & Waibel, A. (2000, October). Application of LDA to speaker recognition. In *Interspeech* (pp. 250-253).
- [38] Mudde, R. F., Groen, J. S., & Van Den Akker, H. E. A. (1998). Application of LDA to bubbly flows. *Nuclear Engineering and Design*, 184(2-3), 329-338

# Identification of Coronary Heart Disease through Iris using Gray Level Co-occurrence Matrix and Support Vector Machine Classification

Vincentius Abdi Gunawan<sup>1\*</sup>  
Department of Informatics Engineering  
Universitas Palangka Raya  
Palangka Raya, Indonesia

Leonardus Sandy Ade Putra<sup>2</sup>, Fitri Imansyah<sup>3</sup>, Eka  
Kusumawardhani<sup>4</sup>  
Department of Electrical Engineering  
Universitas Tanjungpura  
Pontianak, Indonesia

**Abstract**—Now-a-days, coronary heart disease is one of the deadliest diseases in the world. An unfavorable lifestyle, lack of physical activity, and consuming tobacco are the causes of coronary heart disease aside from genetic inheritance. Sometimes the patient does not know whether he has abnormalities in heart function or not. Therefore, this study proposes a system that can detect heart abnormalities through the iris, known as the Iridology method. The system is designed automatically in the iris detection to the classification results. Feature extraction using five characteristics is applied to the Gray Level Co-occurrence Matrix (GLCM) method. The classification process uses the Support Vector Machine (SVM) with linear kernel variation, Polynomial, and Gaussian to obtain the best accuracy in the system. From the system simulation results, the use of the Gaussian kernel can be relied on in the classification of iris conditions with an accuracy rate of 91%, then the Polynomial kernel accuracy reaches 89%, and the linear kernel accuracy reaches 87%. This study has succeeded in detecting heart conditions through the iris by dividing the iris into normal iris and abnormal iris.

**Keywords**—Iris; iridology; coronary heart; circle hough transform; gray level co-occurrence matrix; support vector machine

## I. INTRODUCTION

Coronary heart disease is the number one cause of death worldwide. According to data from the World Health Organization (WHO), there are 17 million people in the world who die from coronary heart disease. In Indonesia, coronary heart disease is the highest cause of death after stroke, with a mortality rate of 12.9% in 2014 [1]. Every year there are 1.9 million people die of coronary heart disease due to consuming tobacco [2]. An unhealthy lifestyle and lack of physical activity are the leading causes of coronary heart disease [3]. The death rate is higher among the older age population [4]. Consuming foods high in carbohydrates or fat and obesity are factors that cause constriction of blood vessels in the heart.

Examination to determine coronary heart disease will be advised by checking with an Electrocardiogram (EKG), which uses electricity to determine heart rhythm. The use of echocardiography is also sometimes done to see the part, pump function, and valve function of the heart. Taking some

actions to check the current heart condition costs quite a bit so that a person will be reluctant to examine his heart condition.

Early prevention can be done to reduce the risk of coronary heart disease, namely by consuming enough fruits and vegetables every day [5], exercising or doing physical activities regularly in daily life, and do not consume tobacco [1].

Nowadays, technology to detect heart conditions from an early stage has been carried out by scientists in the medical field and computing technology. One of them is knowing the heart condition through the eye's iris, which is directly connected to the brain [6]. The brain is a human organ that receives 15% of blood flow from the heart and accounts for 20% of oxygen consumption in the body, making it susceptible to vascularization in the human brain [7].

The iris is one of the unique organs in the human body. Iris is usually used as electronic security or biometric identification system [8,9]. Nowadays, the research on the iris in the medical field is increasingly widespread and is being seriously studied by experts. Iris can provide information about the condition of human organs, known as the Iridology method.

Iridology is based on the analysis of one of the most complex tissue structures contained in the iris. This method can determine the condition of organs and systems in the body from the marks on the iris. Iridology cannot diagnose a disease but instead helps to identify existing or any potential problems in a particular organ [10]. According to Iridologists, the condition of the overall organ is reflected on the surface of the iris [11]. In healthcare practitioners, the iris is used to determine the systemic health, innate strengths, or weaknesses of an individual's personality [12]. Image processing and data mining processing techniques have been used as disease diagnosis tools in biomedical applications. There are parameters in improving the quality of the iris diagnosis technique, especially by using iris images by focusing on the PRISMA-ScR guidelines [13]. Based on this statement, the research obtained 89.63% classification accuracy of diabetic patients with iris diagnosis [14]. Heydari M. and Teimouri M. have also conducted a test on type 2 diabetic patients and obtained the highest accuracy rate of 97.44% using Artificial

\*Corresponding Author

Neural Networks in Iran. Young-WoolLim and Young-BaelPark [11] have also conducted a study to examine the relationship between the iris and the characteristics differences of each individual. The results show that iris parameters can be used more definitely related to characteristics than functional changes. Research conducted by M. Gopalan and Gopal S. Pollai stated that iridology could be used to determine the health of humans organs contained in the character of the iris [12], Recognition of the iris identifies biometrics and can diagnose the presence of cholesterol in the blood [15]. Iris is used in providing information of human's physical condition to determine vehicle driver fatigue by detecting the behavior of the iris that has been carried out by [16], with an accuracy of 80%. Research by [17] has diagnosed the heart with the iris using Principal Component Analysis (PCA) and has achieved 80% success using the SVM classification. The use of the GLCM method was also carried out by [18] by using four characteristics of feature extraction with an accuracy rate of 85.6%.

Previous studies have used the iris as an early diagnosis in detecting abnormal conditions in human organs. The use of different methods leads to different results in identification accuracy. The GLCM method using six characteristics of feature extraction has not been carried out until now. So this is a new scope for in-depth research to be carried out in the identification process. Separating the iris from the rest of the

eye is carried out automatically using CHT by utilizing circle edge detection, an important part of the identification system designed in this paper.

In this study, iris samples were collected from healthy individuals and patients with coronary heart disease to explore new directions and methodologies in diagnosing coronary heart disease with computer imaging and machine learning techniques. The determination of the iris section has been carried out automatically using the Hough transformation method, which is considered more efficient in reading multiple iris samples [19]. Gray Level Co-occurrence Matrix as a texture extraction method was chosen in this study because of its ability to perform feature extraction in grayscale images properly [20]. The selection of GLCM as a feature extraction method is based on previous research that has not tested the condition of the heart through the iris using the six characteristics of GLCM. The GLCM method is considered to be better at reading images with grayscale values to obtain accurate results. Determination of the iris that reflects normal and abnormal heart conditions is carried out using the Support Vector Machine as a classification method. The SVM method can separate data into classes with linear solutions in high dimensions [21]. So that the final results of this study can help provide information on normal heart conditions and heart conditions with abnormal conditions through the iridology method. The system design flow can be shown in Fig. 1.

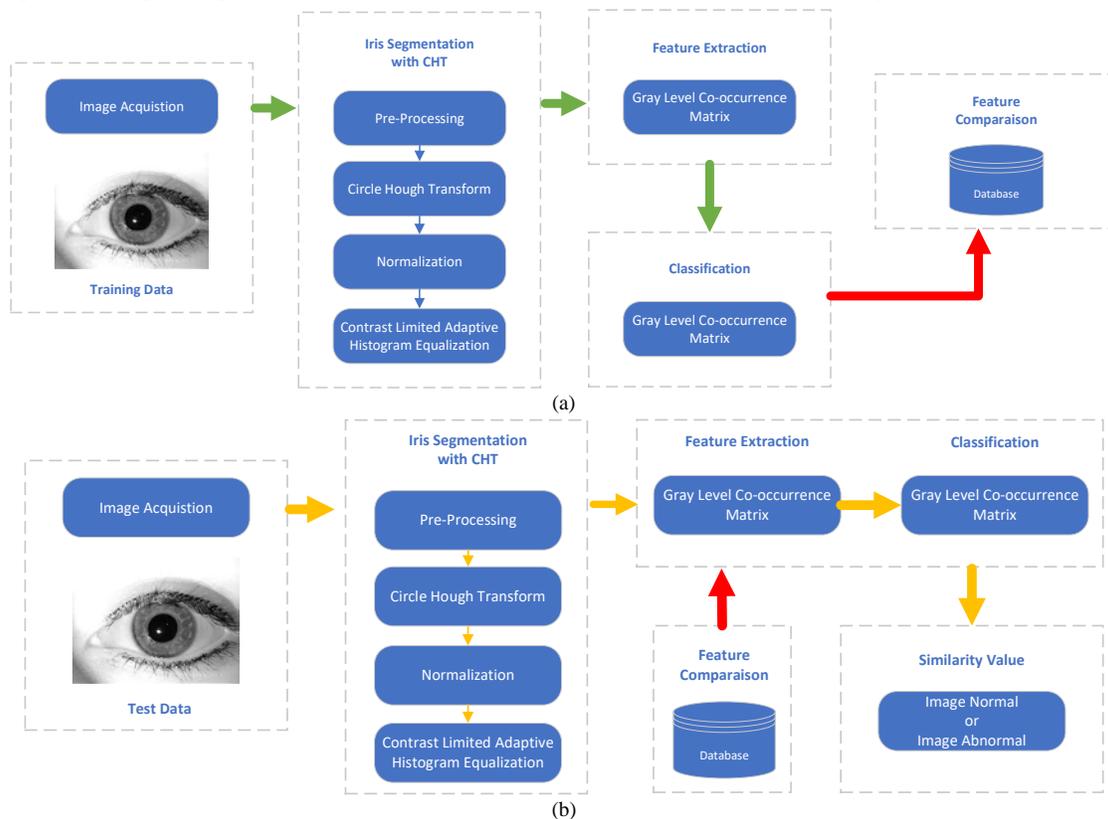


Fig. 1. Shows the Flow of the Designed System; a) the flow of Training on the Recognition System; b) Test Flow on the Recognition System.

## II. RESEARCH METHOD

### A. Iridology

Iridology is known as a diagnostic method using the human iris. In the medical world, the iris can interpret the condition of the human organs [22]. Iridology divides the iris into 60 parts, and each part represents the condition or function of its organs. The right iris will reflect the condition of the right organs of the body, and vice versa, where the left iris will reflect the condition of the left organs [23].

Dr. Bernard Jensen has created a chart that describes each part of the 60 sectors into an image that is mirrored in a circular image like a clock and is divided into sectors according to the part of the iris that reflects the organ. The chart has been described according to the division in the left eye and right eye [10], [24].

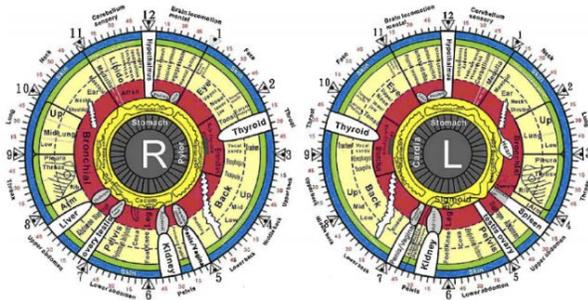


Fig. 2. Left Iris and Right Iris Iridology Chart [10].

We can see in Fig. 2, the position of the heart is on the left side of the iris. The location of the heart is reflected through the left iris, which is shown in the iris zone 02.00 – 03.15.

### B. Pre-Processing

The image captured through a digital camera has an RGB (Red, Green, Blue) format, so it is necessary to do a pre-processing process. In the preprocessing the iris image will be converted from an RGB image to a Greyscale image. Grayscale images have a simpler color value with a color intensity of 0 – 255 pixel thus shortening the computation time. The next step is to cut the unused part outside the iris, leaving a part that is close to the iris. Then, The eye image will

be resized so that the entire image has the same pixel size when it enters the extraction process [25].

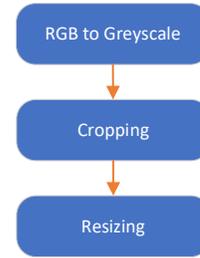


Fig. 3. Image Pre-Processing Process.

Fig. 3 shows the pre-processing chart. The resized image will be processed into the CHT method to obtain the iris and eliminate the pupil area.

### C. Circle Hough Transform

Images that have gone through the RGB to Grayscale conversion process will be separated between the iris and other objects that are not used, especially the pupil. The CHT method is used to determine the iris part automatically without human assistance in determining the coordinates of the iris. CHT can detect circles in the iris image and know between the outer iris and the outer pupil.

$$(x - a)^2 + (y - b)^2 = r^2 \quad (1)$$

Equation 1 describes the center circle  $(a, b)$  and has a radius of  $r$ . With  $(x, y)$  is a pixel at the edge of the circle, en it can be represented in the form of a circle as [26]:

$$\begin{cases} x = a + r \cos \theta \\ y = b + r \sin \theta \end{cases} \quad (2)$$

Fig. 4 describes the CHT method that is implemented into the iris image. The iris image at point (a) has RGB format, which needs to be converted into a greyscale format. The histogram in the grayscale image shows high values in several white areas of objects that occur from light reflections and black colors that have values from the middle pupil of the iris.

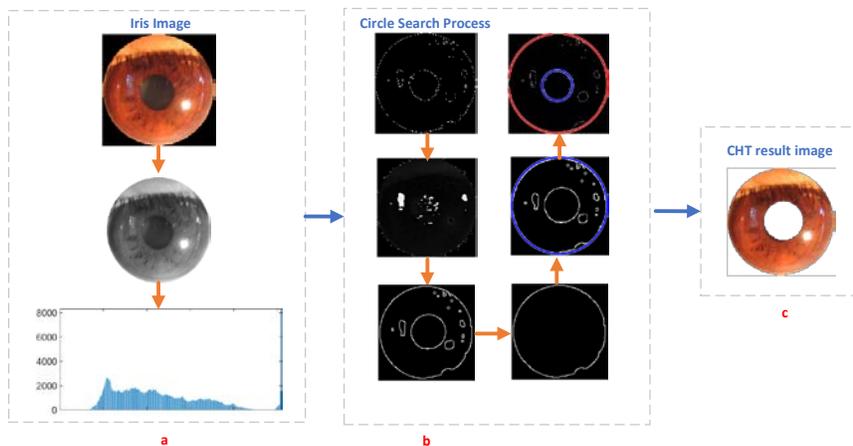


Fig. 4. CHT Process in Detecting Circles on the Iris; a) Iris Image in RGB Format is Converted into Grayscale Format along with Histogram Values; b) Circle Detection Process with CHT; c) the Results of the CHT Process.

Then the image at point b is an image display in determining the point of the circumference of the pupil and iris ring. If a part of the noise is not circular in this process, it will be removed. The image is then given a thickening process on the line to determine the circle in the image. Furthermore, the part that has been detected as a circle will be combined with the initial RGB image and generate an image like point c.

The CHT algorithm is used to separate the iris from the pupil and sclera. This automatic determination is carried out using edge detection in the form of a circle with a diameter value to determine the inner circle and outer circle of the iris. Edge detection is carried out to find objects with a diameter of less than 3 mm for the object removal process. Then the system will detect a circle with a diameter of more than 3 mm with a shape close to a perfect circle that is selected as the inner iris circle. In determining the outline of the outer iris circle, the system looks for the diameter of the circle measuring 12 mm. Determination of the diameter of the circle is adjusted to the process of taking pictures using constant light so that the images in each data taken to have the same size.

#### D. Normalization

The texture in the image has coordinates that represent the dimensions of the iris image, such as pupil dilation [27]. The iris image segmentation method aims to normalize the image in a different form but still has the exact resolution [28]. The iris can be modeled by two non-concentric circles and different textures within an iris circle. The center of the pupil can be used as a reference point for the circle on the iris [8]. A radial line runs through the area in the iris, known as radial resolution. Since the pupil is elastic to the iris, it is necessary to rescaling the reference point. The scaling equation is calculated based on the angle around the iris circle, with the equation:

$$r' = \sqrt{a\beta^2 - a - r_i^2} + \sqrt{a\beta} \quad (3)$$

where,

$$a = o_x^2 + o_y^2 \quad (4)$$

$$\beta = \cos(\pi - \arctan\left(\frac{o_y}{o_x}\right) - \theta) \quad (5)$$

$r'$  is the distance between the pupil and the iris, while  $\theta$  is the edge angle based on the radius of the iris.  $o_x$  and  $o_y$  is the displacement from the center of the pupil to the displacement of the center of the iris [29].

The Iris image that has a circular shape needs to be normalized based on the angle. The circular iris will be formed into a 2D array with horizontal dimensions at angles and vertical dimensions at radial. The circular iris will be formed into a 2D array with horizontal dimensions at angles and vertical dimensions at radial so that it will produce a rectangular or polar image shape as shown in Fig. 5 [30, 31].

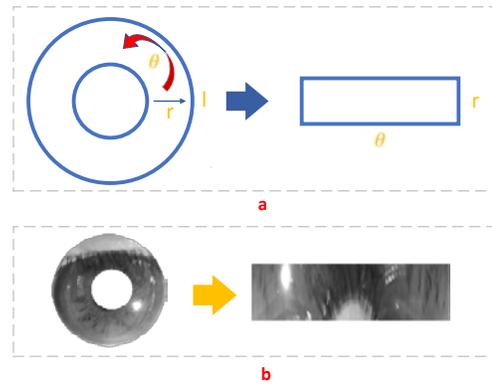


Fig. 5. Area of Circle to Rectangle Transformation of Daugman Model; a) Daugman Model; b) The Implementation Result of Iris Image.

#### E. Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE is a method to improve image quality by limiting the histogram value [32]. In this study, the CLAHE method is used to increase image intensity so that there will be a lot of detail that can be improved on the image. The results of using CLAHE are shown in Fig. 6.

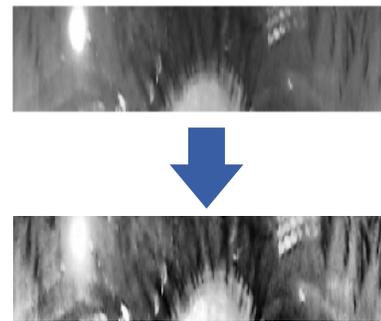


Fig. 6. Results of using the CLAHE Method on Polar Iris Images.

#### F. Region of Interest

The process of diagnosing the heart through the iris requires certain parts so that it does not require the whole to be processed. According to Iridology, the iris that reflects the heart is on the left iris, as shown in the sector 02.00 – 03.15 according to Fig. 1.

Region of Interest (ROI) is a process where we only need a specific part of the image to be processed. These sections will be cropped, leaving only the heart's reflection in the iris of the left, as shown in Fig. 7.

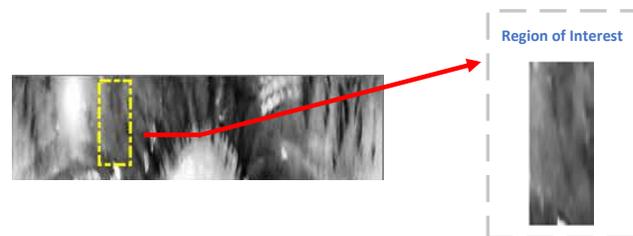


Fig. 7. Application of ROI of the Heart on the Iris Image.

G. Gray Level Co-occurrence Matrix (GLCM)

GLCM is a texture analysis technique on images with a gray level. GLCM has a relationship between 2 neighboring pixels, which is determined by the intensity of gray, a certain distance, and angle. The equation for GLCM is shown below [33]:

$$G_{(\Delta x, \Delta y)}(a, b) = \sum_{i=1}^P \sum_{j=1}^Q 1\{I(i, j) = a\} \text{ and } 1\{I(+\Delta x, j + \Delta y) = b\} \tag{6}$$

$I(i, j)$  is the gray value of the column (i) and row (j) pixels,  $(a, b)$  is a gray value that occurs at the same time as the calculation of  $G_{(\Delta x, \Delta y)}(a, b)$ . Then,  $1\{I(+\Delta x, j + \Delta y) = b\}$  is the indicator of  $\Delta x$  as a direction from  $x$  and  $\Delta y$  as a direction of  $y$  which is determined by the distance of the  $x$  and  $y$ .  $P$  and  $Q$  shows the rows and columns of the corresponding image. There are 4 angles used in GLCM,  $0^\circ, 45^\circ, 90^\circ$  and  $135^\circ$ , described in Fig. 8.

The illustration in Fig. 9 is the use of GLCM for image pixels. Point a shows the use of distances that have a value of  $d = 1$  with 4 different directions where  $d$  is the distance between pixels. Point b shows the GLCM Usage calculation on an image with  $\theta = 0^\circ$  and the distance between pixels is 1.

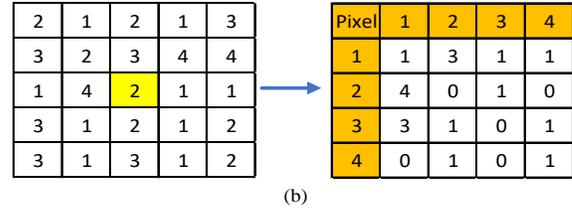
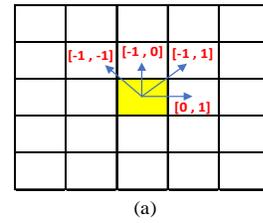


Fig. 9. GLCM Angle Illustration; a) is an Illustration of the use of  $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$  dan  $d = 1$ ; b) GLCM Usage Calculations for Images with  $\theta = 0^\circ$  and  $d = 1$ .

The image that has been calculated for certain distances and angles will then be transposed to the values obtained, and then the GLCM matrix values are added to the transpose results. The results will be normalized using the following equation:

$$GLCM_{Norm} = \frac{GLCM_{value}}{\sum_i^N GLCM_{value}} \tag{7}$$

where,

$GLCM_{value}$  = value of each pixel

The normalization results can be used to determine the texture characteristics of the image by obtaining information, such as Contrast, Dissimilarity, Homogeneity, Angular Second Moment (ASM), Energy and Correlation. The function of the contrast characteristic is the calculation of the difference in intensity between adjacent pixels in the entire image. Dissimilarity is the process of measuring the difference in a texture with a significant value if it is random, and vice versa will have a small value if it is uniform. Homogeneity functions to show the homogeneity of intensity variations in the image. ASM is a uniformity measurement process that produces a high value if the pixel values are similar to each other and vice versa will have a low value if the pixel values are different. Energy is used to measure the concentration of intensity pairs in the matrix, and Correlation is a linearity measurement of several pixel pairs. The equation to obtain each characteristic is shown in Table I [34]:

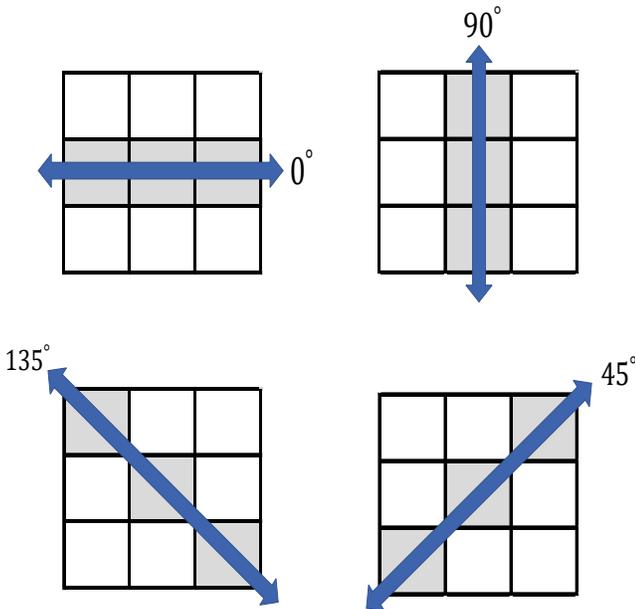


Fig. 8. Illustration of 4-Way Angle GLCM.

TABLE I. CHARACTERISTIC EQUATIONS ON GLCM

| Texture Characteristics of GLCM | Equation                                                                                                                                                                                                                                                                     |
|---------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Contrast                        | $\sum_{a,b=0}^{level-1} P_{a,b}(a-b)^2$                                                                                                                                                                                                                                      |
| Dissimilarity                   | $\sum_{a,b=0}^{level-1} P_{a,b} a-b $                                                                                                                                                                                                                                        |
| Homogeneity                     | $\sum_{a,b=0}^{level-1} \frac{P_{a,b}}{1+(a-b)^2}$                                                                                                                                                                                                                           |
| ASM                             | $\sum_{a,b=0}^{level-1} P_{a,b}^2$                                                                                                                                                                                                                                           |
| Energy                          | $\sqrt{\sum_{a,b=0}^{level-1} P_{a,b}^2}$                                                                                                                                                                                                                                    |
| Correlation                     | $\sum_{a,b=0}^{level-1} P_{a,b} \left[ \frac{(a-\mu_a)(b-\mu_b)}{\sqrt{(\sigma_a^2)(\sigma_b^2)}} \right]$ $\mu_a = \sum_a a \sum_b P_{ab}$ $\mu_b = \sum_b b \sum_a P_{ab}$ $\sigma_a^2 = \sum_a (a-\mu_a)^2 \sum_b P_{ab}$ $\sigma_b^2 = \sum_b (b-\mu_b)^2 \sum_a P_{ab}$ |

where,

$a, b$  = Pixel coordinates on the matrix

$level$  = The range of grayscale value between 0-255 ( $level = 256$ )

$P_{a,b}$  = value of coordinat pixel  $a, b$  on matrix GLCM

#### H. Support Vector Machine

Support Vector machine is one of the supervised machine learning that is commonly used in classification and regression analysis. SVM was widely introduced by Vapnik in the late 90s [35]. SVM can separate data by forming a hyperplane, maximizing margins and dividing data into different classes. The kernel function in SVM can separate data by looking for a hyperplane as a separator between data and finding the best hyperplane to put data in each class [36].

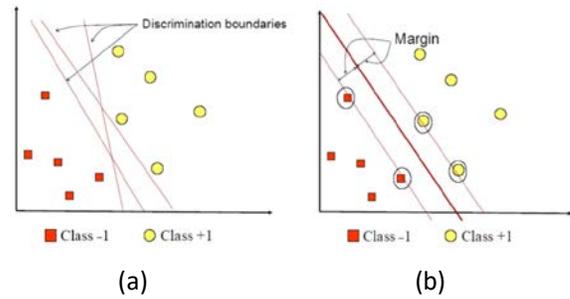


Fig. 10. Illustration of Class -1 and Class -2 Separated by Hyperplane; a) Several Alternative Lines of Discrimination; b) Hyperplane with the Best Margin.

Fig. 10 shows two different classes between class -1 in red and class +1 in yellow. Fig. 5(a) shows classes -1 and +1 separated by several hyperplanes as differentiator classes. The difference can be obtained by measuring the hyperplane margin and finding the maximum point. Margin is the distance between hyperplanes with the closest pattern of each data class. The pattern that has the closest value is also called the support vector. Fig. 5(b) is two classes of data separated by the best hyperplane, where the hyperplane line is located in the middle between the support vector pattern owned by class -1 and class +1.

SVM as a classification method has two essential aspects: the first aspect is finding a hyperplane that can optimally separate two classes. The second aspect is the transformation of linearly inseparable classifications into separable ones [35].

Vector learning input and class  $(x_i, y_i)$ , where  $i = 1, 2, \dots$  with  $x_i \in R^n$  and  $y \in \{1, -1\}$ . In solving problems using linear separation, hyperplane can define boundaries between classes -1 and +1 with the calculation representation as follows. [37]:

$$\min_{\omega, b, \xi} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^l \xi_i \tag{8}$$

towards subject,

$$y_i(\omega^T \phi(x_i) + b) \geq 1 - \xi_i; \xi_i \geq 0 \tag{9}$$

Solving for classes that cannot be solved linearly, then vector  $x_i$  mapped into a higher dimension using the function  $\phi$  so that they can be separated linearly. Furthermore, SVM can define a hyperplane that separates linearly with a maximum margin of higher dimensional space. The SVM error parameter is known as  $C > 0$ .  $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$  known as the Kernel function, shown in Table II [38].

TABLE II. SVM KERNEL EQUATION

| Kernel Name | Equation                                                                       |
|-------------|--------------------------------------------------------------------------------|
| Linear      | $K(x_i, x_j) = x_i^T x_j$                                                      |
| Polynomial  | $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$                           |
| Gaussian    | $K(x_i, x_j) = \exp\left(-\frac{\ x_i - x_j\ ^2}{2\gamma^2}\right) \gamma > 0$ |

### III. RESULTS AND ANALYSIS

In this study, system training was carried out using 75 normal iris data and 75 abnormal iris data. Normal iris data is the iris of people who have no history of heart disease; on the contrary for abnormal iris data is the iris of people who have heart disease.

Fig. 11 shows the training data using linear, polynomial, and gaussian kernel variations. Iris data in training can be separated according to normal (red) and abnormal (blue) classes. The results of linear kernel training separate the data into each class with an even distribution of data. The difference in the polynomial kernel training where the data has been separated and more centralized. The results of the training using the Gaussian Kernel resulted in a tighter grouping than using the two previous kernels. The results of the training are able to separate between classes according to existing characteristics, which can help in the classification of test data and affect the level of recognition accuracy. The further apart the hyperplane in the SVM that separates the classes, the higher the accuracy.

Tests were carried out on 50 normal iris data and 50 abnormal iris data. The data that has gone through the region of interest process is the iris data which reflects the heart. Then this section will be extracted using the GLCM method. The characteristics used in GLCM are Contrast, Dissimilarity,

Homogeneity, Energy, and Correlation. The use of angle variations in the test is carried out to obtain optimal results. For example, Fig. 12 results from GLCM feature extraction at an angle with a distance of 1.

Taking five types of texture characteristics on iris data is needed to obtain more detailed information in training and testing. Normal and abnormal iris data have different characteristic values. The data can be classified into the appropriate class classification using SVM with three kernels: Linear, Polynomial, and Gaussian.

Simulation testing for variations angle in GLCM  $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$  with the SVM kernel variation is shown in Fig. 13. By using angle  $0^\circ$  the highest level of accuracy is obtained by using the Gaussian kernel which reaches 94%, linear kernel reaches 90% and Polynomial kernel 88%. The use of  $45^\circ$  obtained the highest accuracy rate with 88% Gaussian kernel, 86% polynomial kernel and 82% linear kernel. Then, for angel  $90^\circ$  the highest level of accuracy is achieved by Gaussian and Linear kernels which reach 92% and polynomial kernels which reach 88%. Meanwhile, with angle  $135^\circ$  the highest level of accuracy is achieved by 92% polynomial kernel, 90% Gaussian kernel and linear kernel which reaches 84%. From these data, it can be seen that the angle  $0^\circ$  and  $90^\circ$  has an average classification accuracy level of 90.67% at a distance of 1.

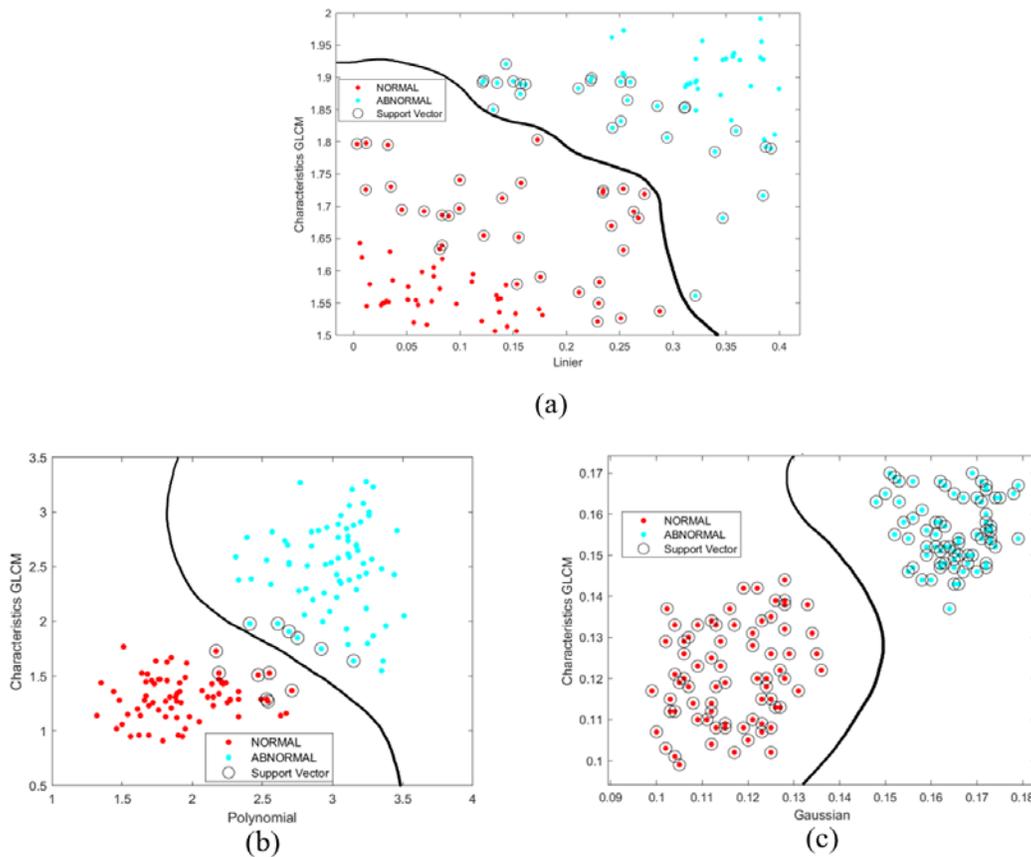


Fig. 11. Results of Training using Kernel Variations; a) Linear kernel; b) Polynomial kernel; c) Gaussian kernel.

| Iris Name | Contrast | Dissimilarity | Homogeneity | Energy   | Correlation | Output Target |
|-----------|----------|---------------|-------------|----------|-------------|---------------|
| iris_1_N  | 0.113840 | 1.534803      | 0.875909    | 0.223541 | 0.530184    | Normal        |
| iris_2_N  | 0.137012 | 1.556464      | 0.909280    | 0.242990 | 0.567106    | Normal        |
| iris_3_N  | 0.092625 | 1.514334      | 0.861357    | 0.205152 | 0.509539    | Normal        |
| iris_4_N  | 0.126065 | 1.547119      | 0.890755    | 0.230201 | 0.536350    | Normal        |
| iris_5_N  | 0.105937 | 1.530152      | 0.869887    | 0.215905 | 0.521832    | Normal        |
| iris_6_N  | 0.131387 | 1.554229      | 0.904014    | 0.237140 | 0.561935    | Normal        |
| iris_7_N  | 0.108102 | 1.533261      | 0.871327    | 0.217578 | 0.522959    | Normal        |
| iris_8_N  | 0.108949 | 1.533432      | 0.871362    | 0.218346 | 0.523537    | Normal        |
| iris_9_N  | 0.120500 | 1.542982      | 0.884436    | 0.227554 | 0.532651    | Normal        |
| iris_10_N | 0.114233 | 1.536698      | 0.876382    | 0.224434 | 0.530588    | Normal        |
| iris_11_N | 0.122934 | 1.545506      | 0.888404    | 0.229146 | 0.533992    | Normal        |
| iris_12_N | 0.117435 | 1.540503      | 0.882702    | 0.225038 | 0.531913    | Normal        |
| iris_13_N | 0.099871 | 1.525741      | 0.866672    | 0.214070 | 0.518462    | Normal        |
| iris_14_N | 0.095196 | 1.518999      | 0.862712    | 0.206916 | 0.513388    | Normal        |
| iris_15_N | 0.130812 | 1.553009      | 0.901850    | 0.236055 | 0.557268    | Normal        |
| iris_16_N | 0.091719 | 1.511600      | 0.801879    | 0.200253 | 0.505685    | Normal        |
| iris_17_N | 0.129111 | 1.550672      | 0.900261    | 0.232182 | 0.544913    | Normal        |
| iris_18_N | 0.135080 | 1.555202      | 0.908694    | 0.240291 | 0.565829    | Normal        |
| iris_19_N | 0.099355 | 1.524967      | 0.865813    | 0.208856 | 0.516272    | Normal        |
| iris_20_N | 0.122329 | 1.544525      | 0.888029    | 0.228287 | 0.533974    | Normal        |
| iris_21_N | 0.128799 | 1.550137      | 0.893426    | 0.231181 | 0.543751    | Normal        |
| iris_22_N | 0.106555 | 1.532400      | 0.870643    | 0.216479 | 0.522577    | Normal        |
| iris_23_N | 0.121230 | 1.544125      | 0.885671    | 0.227872 | 0.532727    | Normal        |
| iris_24_N | 0.113138 | 1.534629      | 0.874729    | 0.222447 | 0.530116    | Normal        |
| iris_25_N | 0.117625 | 1.541798      | 0.882963    | 0.225234 | 0.532239    | Normal        |
| iris_26_N | 0.102282 | 1.527922      | 0.868529    | 0.214469 | 0.519255    | Normal        |
| iris_27_N | 0.099305 | 1.521108      | 0.865209    | 0.207950 | 0.514743    | Normal        |
| iris_28_N | 0.112204 | 1.533513      | 0.872597    | 0.219700 | 0.524008    | Normal        |
| iris_29_N | 0.138300 | 1.556943      | 0.909989    | 0.244259 | 0.567959    | Normal        |
| iris_30_N | 0.114878 | 1.537424      | 0.877369    | 0.224701 | 0.530918    | Normal        |
| iris_31_N | 0.127546 | 1.550092      | 0.893227    | 0.230965 | 0.542584    | Normal        |
| iris_32_N | 0.093969 | 1.515478      | 0.862557    | 0.205657 | 0.512927    | Normal        |
| iris_33_N | 0.132381 | 1.554773      | 0.905114    | 0.238878 | 0.563652    | Normal        |
| iris_34_N | 0.126967 | 1.548483      | 0.891537    | 0.230536 | 0.539478    | Normal        |
| iris_35_N | 0.115360 | 1.540278      | 0.877390    | 0.224781 | 0.531348    | Normal        |
| iris_36_N | 0.112290 | 1.534626      | 0.872953    | 0.220016 | 0.529340    | Normal        |
| iris_37_N | 0.129724 | 1.551908      | 0.900730    | 0.235930 | 0.549226    | Normal        |
| iris_38_N | 0.135673 | 1.555577      | 0.908772    | 0.241097 | 0.566311    | Normal        |
| iris_39_N | 0.119169 | 1.542638      | 0.883913    | 0.227447 | 0.532550    | Normal        |
| iris_40_N | 0.092119 | 1.511946      | 0.804009    | 0.202494 | 0.505771    | Normal        |
| iris_41_N | 0.118351 | 1.541898      | 0.883264    | 0.227327 | 0.532538    | Normal        |
| iris_42_N | 0.125965 | 1.547022      | 0.889048    | 0.229819 | 0.536237    | Normal        |
| iris_43_N | 0.135992 | 1.556369      | 0.909058    | 0.242795 | 0.566957    | Normal        |
| iris_44_N | 0.103427 | 1.529705      | 0.869788    | 0.214554 | 0.521212    | Normal        |
| iris_45_N | 0.131376 | 1.554161      | 0.903425    | 0.237118 | 0.561623    | Normal        |
| iris_46_N | 0.096582 | 1.519643      | 0.863847    | 0.207312 | 0.514044    | Normal        |
| iris_47_N | 0.097895 | 1.520501      | 0.864548    | 0.207614 | 0.514577    | Normal        |
| iris_48_N | 0.094387 | 1.518236      | 0.862694    | 0.205886 | 0.513336    | Normal        |
| iris_49_N | 0.131368 | 1.553203      | 0.903417    | 0.237117 | 0.557788    | Normal        |
| iris_50_N | 0.099767 | 1.525568      | 0.865828    | 0.210194 | 0.517831    | Normal        |

| Iris Name  | Contrast | Dissimilarity | Homogeneity | Energy   | Correlation | Output Target |
|------------|----------|---------------|-------------|----------|-------------|---------------|
| iris_1_AB  | 0.329457 | 1.906424      | 0.586281    | 0.437836 | 0.834653    | Abnormal      |
| iris_2_AB  | 0.357422 | 1.936454      | 0.612558    | 0.457730 | 0.860977    | Abnormal      |
| iris_3_AB  | 0.311232 | 1.883080      | 0.571065    | 0.415062 | 0.810081    | Abnormal      |
| iris_4_AB  | 0.343302 | 1.921688      | 0.603311    | 0.447064 | 0.849877    | Abnormal      |
| iris_5_AB  | 0.322849 | 1.895183      | 0.581332    | 0.433354 | 0.823621    | Abnormal      |
| iris_6_AB  | 0.351030 | 1.930470      | 0.607544    | 0.450543 | 0.858039    | Abnormal      |
| iris_7_AB  | 0.323676 | 1.898782      | 0.582187    | 0.434065 | 0.825033    | Abnormal      |
| iris_8_AB  | 0.325459 | 1.903395      | 0.583222    | 0.434778 | 0.827332    | Abnormal      |
| iris_9_AB  | 0.337747 | 1.917268      | 0.599114    | 0.441101 | 0.844149    | Abnormal      |
| iris_10_AB | 0.331364 | 1.906475      | 0.586905    | 0.438649 | 0.835570    | Abnormal      |
| iris_11_AB | 0.342754 | 1.919736      | 0.602187    | 0.444860 | 0.848469    | Abnormal      |
| iris_12_AB | 0.333855 | 1.915622      | 0.587599    | 0.438992 | 0.839845    | Abnormal      |
| iris_13_AB | 0.321665 | 1.892886      | 0.578839    | 0.430163 | 0.820626    | Abnormal      |
| iris_14_AB | 0.312733 | 1.886257      | 0.574593    | 0.422003 | 0.814229    | Abnormal      |
| iris_15_AB | 0.348505 | 1.927886      | 0.607177    | 0.449208 | 0.856424    | Abnormal      |
| iris_16_AB | 0.309968 | 1.882335      | 0.505143    | 0.413547 | 0.807357    | Abnormal      |
| iris_17_AB | 0.346304 | 1.926923      | 0.604156    | 0.448388 | 0.856050    | Abnormal      |
| iris_18_AB | 0.354216 | 1.932049      | 0.609695    | 0.454194 | 0.858625    | Abnormal      |
| iris_19_AB | 0.321182 | 1.891907      | 0.578466    | 0.429390 | 0.818582    | Abnormal      |
| iris_20_AB | 0.340548 | 1.918924      | 0.600069    | 0.444651 | 0.847406    | Abnormal      |
| iris_21_AB | 0.346011 | 1.924694      | 0.604131    | 0.447560 | 0.854405    | Abnormal      |
| iris_22_AB | 0.323357 | 1.896666      | 0.581541    | 0.434037 | 0.824947    | Abnormal      |
| iris_23_AB | 0.339513 | 1.918457      | 0.599624    | 0.442365 | 0.844884    | Abnormal      |
| iris_24_AB | 0.328327 | 1.905546      | 0.585306    | 0.436282 | 0.834512    | Abnormal      |
| iris_25_AB | 0.334689 | 1.916217      | 0.588875    | 0.439787 | 0.840532    | Abnormal      |
| iris_26_AB | 0.321787 | 1.894896      | 0.578999    | 0.431165 | 0.823337    | Abnormal      |
| iris_27_AB | 0.316530 | 1.890742      | 0.577807    | 0.427436 | 0.817862    | Abnormal      |
| iris_28_AB | 0.325815 | 1.903526      | 0.583535    | 0.436119 | 0.829093    | Abnormal      |
| iris_29_AB | 0.357517 | 1.937150      | 0.614173    | 0.458252 | 0.861316    | Abnormal      |
| iris_30_AB | 0.332150 | 1.911090      | 0.587349    | 0.438668 | 0.838154    | Abnormal      |
| iris_31_AB | 0.344889 | 1.924101      | 0.603769    | 0.447470 | 0.853389    | Abnormal      |
| iris_32_AB | 0.311703 | 1.883815      | 0.571998    | 0.419483 | 0.813329    | Abnormal      |
| iris_33_AB | 0.351180 | 1.931843      | 0.608249    | 0.452731 | 0.858070    | Abnormal      |
| iris_34_AB | 0.343480 | 1.922127      | 0.603341    | 0.447468 | 0.850663    | Abnormal      |
| iris_35_AB | 0.333213 | 1.913168      | 0.587382    | 0.438902 | 0.838449    | Abnormal      |
| iris_36_AB | 0.327655 | 1.903563      | 0.585244    | 0.436248 | 0.833949    | Abnormal      |
| iris_37_AB | 0.346971 | 1.927239      | 0.605717    | 0.448549 | 0.856358    | Abnormal      |
| iris_38_AB | 0.355019 | 1.933204      | 0.609905    | 0.454372 | 0.859139    | Abnormal      |
| iris_39_AB | 0.335945 | 1.916977      | 0.593943    | 0.440565 | 0.843881    | Abnormal      |
| iris_40_AB | 0.310655 | 1.882698      | 0.568374    | 0.414822 | 0.807485    | Abnormal      |
| iris_41_AB | 0.334799 | 1.916866      | 0.589686    | 0.439879 | 0.842201    | Abnormal      |
| iris_42_AB | 0.343254 | 1.920767      | 0.602435    | 0.447036 | 0.849754    | Abnormal      |
| iris_43_AB | 0.356754 | 1.934325      | 0.611557    | 0.456995 | 0.860948    | Abnormal      |
| iris_44_AB | 0.322293 | 1.894955      | 0.579395    | 0.431568 | 0.823568    | Abnormal      |
| iris_45_AB | 0.350102 | 1.929388      | 0.607501    | 0.450050 | 0.857980    | Abnormal      |
| iris_46_AB | 0.313168 | 1.889213      | 0.575137    | 0.424492 | 0.815853    | Abnormal      |
| iris_47_AB | 0.314682 | 1.890048      | 0.576256    | 0.426567 | 0.817348    | Abnormal      |
| iris_48_AB | 0.311746 | 1.885040      | 0.573339    | 0.419889 | 0.813650    | Abnormal      |
| iris_49_AB | 0.349897 | 1.929132      | 0.607417    | 0.449636 | 0.856578    | Abnormal      |
| iris_50_AB | 0.321245 | 1.892147      | 0.578643    | 0.429447 | 0.819186    | Abnormal      |

Fig. 12. GLCM Characteristic Value using Angle 0° and d = 1 on Normal and Abnormal Iris Data.



Fig. 13. The Results of the Test with Variations of the GLCM Angle on the Variation of the SVM Kernel.

TABLE III. THE AVERAGE LEVEL OF ANGLE ACCURACY ON THE VARIATION OF THE SVM KERNEL

| SVM Classifiers | Training Time (s) | Accuracy Percentage |
|-----------------|-------------------|---------------------|
| Linier          | 381.73            | 87%                 |
| Polynomial      | 422.28            | 89%                 |
| Gaussian        | 583.09            | 91%                 |

From Table III, it can be concluded that the average for each angle variation using different SVM kernels has better performance in the classification process. By testing 100 normal and abnormal iris data, a high degree of accuracy was obtained. The use of linear SVM obtained an accuracy rate of 87%, then the use of polynomials obtained 89%. While the use of the kernel with the highest accuracy reaching 91%, was obtained using Gaussian.

Simulation tests with the Gaussian kernel proved to be superior in iris image classification. The Gaussian kernel

considers the probability of the density function of the standard deviation, squared, and variance. Gaussian can add data space into a higher dimensional vector to determine the intersection of hyperplanes more accurately. With flexible limiting performance, the Gaussian kernel can obtain a higher accuracy level than Linear and Polynomial kernels.

It is proven that normal and abnormal iris conditions can be identified using GLCM feature extraction and a classification process using the Gaussian method on the iris image. Identification is made to assist in providing an early diagnosis of heart conditions through the iris using the iridology method.

#### IV. CONCLUSION

This study has proposed a new method to determine the condition of the heart through the iris using the SVM classification with variations of the linear kernel, polynomial kernel, and Gaussian kernel. The use of GLCM characteristics as feature extraction has an essential role in the classification process. The main contribution in this study is not only limited to determining heart health conditions through the iris but also contributes to the automatic processing of the iris with CHT. The proposed system in determining the heart condition automatically is to optimize the classification by using angle  $0^\circ$  and  $90^\circ$  on GLCM with SVM classification on the Gaussian kernel to obtain a high level of accuracy. In ongoing research, the iris database can be added to improve the classification to make it more accurate. Different extraction methods can be used to get the results of image extraction with a smaller size so that it can increase the system's speed in iris identification.

#### ACKNOWLEDGMENT

This work was supported by the Kariadi Hospital in Semarang and the Department of Engineering at the Universitas Palangkaraya and the Department of Engineering at the Universitas Tanjungpura, Indonesia.

#### REFERENCES

- [1] Kementerian Kesehatan Republik Indonesia, "Hari Jantung Sedunia (HJS) Tahun 2019 : Jantung Sehat, SDM Unggul - Direktorat P2PTM," P2PTM Kemenkes RI. 2019, [Online]. Available: <http://p2ptm.kemkes.go.id/kegiatan-p2ptm/pusat/hari-jantung-sedunia-hjs-tahun-2019-jantung-sehat-sdm-unggul>.
- [2] WHO, "10 Facts on Ageing and Health," Who, vol. 2050, no. May 2017, p. 2014, 2016.
- [3] E. E. Tuppo, M. P. Trivedi, J. B. Kostis, J. Daevmer, J. Cabrera, and W. J. Kostis, "The role of public health versus invasive coronary interventions in the decline of coronary heart disease mortality," *Ann. Epidemiol.*, vol. 55, pp. 91–97, 2021, doi: 10.1016/j.annepidem.2020.10.005.
- [4] C. Murray, C. Atkinson, K. Bhalla, G. Birbeck, R. Burstein, and D. Chou, "The state of US health, 1990-2010: burden of diseases, injuries, and risk factors," *JAMA - J. Am. Med. Assoc.*, vol. 310, no. 6, pp. 591–608, 2013, doi: 10.1001/jama.2013.13805.The.
- [5] Y. H. Chiu et al., "Association between intake of fruits and vegetables by pesticide residue status and coronary heart disease risk," *Environ. Int.*, vol. 132, no. May, p. 105113, 2019, doi: 10.1016/j.envint.2019.105113.
- [6] E. G. Nabel and E. Braunwald, "A Tale of Coronary Artery Disease and Myocardial Infarction," pp. 54–63, 2012.
- [7] F. J. Wolters et al., "Coronary heart disease, heart failure, and the risk of dementia: A systematic review and meta-analysis," *Alzheimer's Dement.*, vol. 14, no. 11, pp. 1493–1504, 2018, doi: 10.1016/j.jalz.2018.01.007.
- [8] R. Biswas, J. Uddin, and M. J. Hasan, "A new approach of iris detection and recognition," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 5, pp. 2530–2536, 2017, doi: 10.11591/ijece.v7i5.pp2530-2536.
- [9] H. Ohmaid, S. Eddarouich, A. Bourouhou, and M. Timouyas, "Iris segmentation using a new unsupervised neural approach," *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, pp. 58–64, 2020, doi: 10.11591/ijai.v9.i1.pp58-64.
- [10] B. Jensen, "Science of Iridology," pp. 1–2, 1982.
- [11] J. Deck, "Principles of Iris Diagnosis," no. June 1985, pp. 1–4.
- [12] M. Gopalan and G. S. Pillai, "Human iris patterns – Iridology – Applications," *J. Anat. Soc. India*, vol. 67, p. S68, Aug. 2018, doi: 10.1016/j.jasi.2018.06.132.
- [13] R. B. Esteves, J. A. P. Morero, S. de S. Pereira, K. D. S. Mendes, K. M. Hegadoren, and L. Cardoso, "Parameters to increase the quality of iridology studies: A scoping review," *Eur. J. Integr. Med.*, vol. 43, p. 101311, Apr. 2021, doi: 10.1016/j.eujim.2021.101311.
- [14] P. Samant and R. Agarwal, "Machine learning techniques for medical diagnosis of diabetes using iris images," *Comput. Methods Programs Biomed.*, vol. 157, pp. 121–128, Apr. 2018, doi: 10.1016/j.cmpb.2018.01.004.
- [15] R. A. Ramlee and S. Ranjit, "Using iris recognition algorithm, detecting cholesterol presence," *Proc. - 2009 Int. Conf. Inf. Manag. Eng. ICIME 2009*, pp. 714–717, 2009, doi: 10.1109/ICIME.2009.61.
- [16] K. Gopalakrishna and S. A. Hariprasad, "Real-time fatigue analysis of driver through iris recognition," *Int. J. Electr. Comput. Eng.*, vol. 7, no. 6, pp. 3306–3312, 2017, doi: 10.11591/ijece.v7i6.pp3306-3312.
- [17] L. I. Permatasari, A. Novianty, and T. W. Purboyo, "Heart disorder detection based on computerized iridology using support vector machine," *ICCEREC 2016 - Int. Conf. Control. Electron. Renew. Energy, Commun. 2016, Conf. Proc.*, pp. 157–161, 2017, doi: 10.1109/ICCEREC.2016.7814983.
- [18] R. Aminah and A. H. Saputro, "Application of machine learning techniques for diagnosis of diabetes based on iridology," *2019 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2019*, pp. 133–138, 2019, doi: 10.1109/ICACSIS47736.2019.8979755.
- [19] I. A. Qasmieh, H. Alquran, and A. M. Alqudah, "Occluded iris classification and segmentation using self-customized artificial intelligence models and iterative randomized Hough transform," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 5, pp. 4037–4049, 2021, doi: 10.11591/ijece.v11i5.pp4037-4049.
- [20] T. Chekouo, S. Mohammed, and A. Rao, "A Bayesian 2D functional linear model for gray-level co-occurrence matrices in texture analysis of lower grade gliomas," *NeuroImage Clin.*, vol. 28, p. 102437, Jan. 2020, doi: 10.1016/j.nicl.2020.102437.
- [21] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, 1998.
- [22] D. C. Adelina, R. Sigit, T. Harsono, and M. Rochmad, "Identification Of Diabetes In Pancreatic Organs Using Iridology," pp. 114–119, 2017.
- [23] L. F. Salles and M. J. P. de Silva, "Iridology: A systematic review," *Rev. da Esc. Enferm.*, vol. 42, no. 3, pp. 585–589, 2008, doi: 21201.
- [24] S. E. Hussein, O. A. Hassan, and M. H. Granat, "Assessment of the potential iridology for diagnosing kidney disease using wavelet analysis and neural networks," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 534–541, 2013, doi: 10.1016/j.bspc.2013.04.006.
- [25] L. S. A. Putra, L. Sumarno, and V. A. Gunawan, "The recognition of semaphore letter code using haar wavelet and euclidean function," *Int. Conf. Electr. Eng. Comput. Sci. Informatics, vol. 2018-October*, no. 1, pp. 759–763, 2018, doi: 10.1109/EECSI.2018.8752707.
- [26] A. O. Djekoune, K. Messaoudi, and K. Amara, "Incremental circle hough transform: An improved method for circle detection," *Optik (Stuttg.)*, vol. 133, pp. 17–31, Mar. 2017, doi: 10.1016/j.ijleo.2016.12.064.
- [27] T. Lefevre, B. Dorizzi, S. Garcia-Salicetti, N. Lemperiere, and S. Belardi, "Effective elliptic fitting for iris normalization," *Comput. Vis. Image Underst.*, vol. 117, no. 6, pp. 732–745, Jun. 2013, doi: 10.1016/j.cviu.2013.01.005.

- [28] J. Daugman, "High Conf Visual Recog of Persons by a test of statistical significance PAMI93," *Ieee Pami*, vol. 15, no. 11, 1993.
- [29] G. Indrawan, S. Akbar, and B. Sitohang, "Fingerprint direct-access strategy using local-star-structure-based discriminator features: A comparison study," *Int. J. Electr. Comput. Eng.*, vol. 4, no. 5, pp. 817–830, 2014, doi: 10.11591/ijece.v4i5.6589.
- [30] L. S. Ade Putra, R. Rizal Isnanto, A. Triwiyatno, and V. A. Gunawan, "Identification of Heart Disease with Iridology Using Backpropagation Neural Network," 2018 2nd Borneo Int. Conf. Appl. Math. Eng. BICAME 2018, pp. 138–142, 2018, doi: 10.1109/BICAME45512.2018.1570509882.
- [31] R. P. Wildes, "Iris recognition: An emerging biometric technology," *Proc. IEEE*, vol. 85, no. 9, pp. 1348–1363, 1997, doi: 10.1109/5.628669.
- [32] R. A. Manju, G. Koshy, and P. Simon, "Improved Method for Enhancing Dark Images based on CLAHE and Morphological Reconstruction," *Procedia Comput. Sci.*, vol. 165, no. 2019, pp. 391–398, 2019, doi: 10.1016/j.procs.2020.01.033.
- [33] V. Naghashi, "Co-occurrence of adjacent sparse local ternary patterns: A feature descriptor for texture and face image retrieval," *Optik (Stuttg.)*, vol. 157, pp. 877–889, Mar. 2018, doi: 10.1016/j.ijleo.2017.11.160.
- [34] D. W. Yang and H. Wu, "Three-dimensional temperature uniformity assessment based on gray level co-occurrence matrix," *Appl. Therm. Eng.*, vol. 108, pp. 689–696, Sep. 2016, doi: 10.1016/j.applthermaleng.2016.07.145.
- [35] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," *Telkomnika (Telecommunication Comput. Electron. Control.)*, vol. 18, no. 2, pp. 815–821, 2020, doi: 10.12928/TELKOMNIKA.V18I2.14785.
- [36] S. Widodo, R. N. Rohmah, B. Handaga, and L. D. D. Arini, "Lung diseases detection caused by smoking using support vector machine," *Telkomnika (Telecommunication Comput. Electron. Control.)*, vol. 17, no. 3, pp. 1256–1266, 2019, doi: 10.12928/TELKOMNIKA.V17I3.9799.
- [37] H. Ohmaid, S. Eddarouich, A. Bourouhou, and M. Timouyas, "Comparison between svm and knn classifiers for iris recognition using a new unsupervised neural approach in segmentation," *IAES Int. J. Artif. Intell.*, vol. 9, no. 3, pp. 429–438, 2020, doi: 10.11591/ijai.v9.i3.pp429-438.
- [38] A. Czajka, K. W. Bowyer, M. Krumdick, and R. G. Vidalmata, "Recognition of Image-Orientation-Based Iris Spoofing," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 9, pp. 2184–2196, 2017, doi: 10.1109/TIFS.2017.2701332.

# Performance of Data Reduction Algorithms for Wireless Sensor Network (WSN) using Different Real-Time Datasets: Analysis Study

M. K. Hussein<sup>1</sup>, Ion Marghescu<sup>2</sup>

Dept. Electronics, Telecommunication & Information  
Technology  
University Politehnica of Bucharest  
Bucharest, Romania

Nayef.A.M. Alduais<sup>3</sup>

Faculty of Computer Science and Information Technology  
Universiti Tun Hussein Onn Malaysia  
Johor, Malaysia

**Abstract**—This paper investigates the effect of data reduction methods in the performance of Wireless Sensor Network (WSN) using a variety of real-time datasets. The simulation tests are carried out in MATLAB for several methods of reducing the quantity of sent data. These approaches are Data Reduction based - Neural Network Fitting (NNF), Neural Network Time Series (NNTS), Linear Regression with Multiple Variables (LRMV), Data Reduction based – “An Efficient Data Collection and Dissemination (EDCD2)” and Data Reduction based – Fast Independent Component Analysis (FICA). The selected algorithms NNF, NNST, EDCD2, LRMV, and FICA are evaluated using real-time datasets. The performance indicators included are energy consumption, data accuracy, and data reduction percentage. The research results show that the selected algorithm helps to reduce the amount of data transferred and consumed energy, but each algorithm performs differently depending on the dataset used.

**Keywords**—Data reduction algorithms; WSN; energy consumption; accuracy; neural network; independent component analysis

## I. INTRODUCTION

In this paper, Wireless Sensor Network (WSN) is a network that collects data from spatially isolated sensors. Sensor nodes are used to monitor and record environmental variables, such as sound, pollution level, humidity, temperature, and wind, and then send the sensed data to the base station [1][2]. The sensor node in the WSNs application is powered by a battery with limited-service life [3]. Furthermore, sensor nodes with multivariable sensors can have an impact on battery life because the node must support additional data transmission, causing the battery to drain faster than a sensor node with a single sensor [4]. Therefore, many researchers have been proposed various approaches to reduce the amount of transmitted data at the sensor node level, which will help in prolonging the battery lifetime. For example, in WSN, the spatial and temporal correlation between the generated traffic can be used to reduce the energy consumption of continuous sensor data acquisition. Spatial-temporal correlation is used in dual prediction (DP) and data compression (DC) techniques to reduce the number of transmissions to save energy and bandwidth. In [5], the author has used these two technologies as part of a two-stage data reduction scheme. The DP

technology reduces traffic between cluster nodes and cluster heads, while the DC scheme reduces traffic between cluster heads and sink nodes.

In [6], the author proposed a data-aware energy-saving technology. The essential correlation between continuous measurements of sensor nodes and the similarity of data trends between adjacent sensor nodes are used to reduce data transmissions. The forecast-based data collection framework reduces time data redundancy. “Autoregressive Integrated Moving Average (ARIMA)” model was used to predict data. The proposed model was implemented in the Cluster head (CH) node.

In [7], the author proposed a novel technique for secure data prediction in WSN by using a Time Series Trust Model (TSTM) based on the Toeplitz matrix and a trust-based autoregressive process (TAR). The author proposed an adaptive data reduction method (AM-DR) in [8]. AM-DR is based on a convex combination of two decoupled Least-Mean-Square (LMS) window filters of different widths for predicting the next readings at both the source and sink nodes.

In [9], the authors have evaluated the performance of several methods based on computational intelligence to decrease the amount of the payload of every packets sent from the sensor node to the base station. These approaches are data reduction based on “artificial neural networks (DR-ANN)”; independent component analysis (DR-ICA) and deep learning regression methods called DR-GDMLR”.

In [10], two multivariate data reduction methods for adaptive thresholds were proposed a Principal Component Analysis Based (PCA-B) –and multiple linear regression Based (MLR-B). PCA-B is a multivariate data reduction method. It uses “Candid Covariance-free Incremental PCA (CCIPCA)” with an adaptive threshold and to reach a high reduction ratio the number of Principal Components (PCs) assigned to “1”. Another method to decrease the amount of payload sensed data is named MLR-B, which it using multiple linear regression (MLR) model. The authors used an adaptive threshold to retrain the model. According to [10], after updating the reference parameters of the model, the size of the transmitted data is greater than or equal to the size of the payload data without being reduced. This means that the sensor node needs

more power during the update phase, than the required power during the phase of reduction. The article recommends a new indicator for evaluating the performance of data reduction models, which it considering the number of repeating of updating the parameters reference of the model. A novel simple scheme called “Adaptive Real-Time Payload Data Reduction Scheme (APRS)” is proposed in [4]. APRS purposes is to decrease the size of the transferred payload of the sensor nodes. Further details on approaches of data reduction for sensor nodes are defined in [11]. In this study, effect of data reduction approaches on WSN performance is investigated using a set of real-time datasets. Simulation tests are performed in MATLAB for different approaches to decrease the amount of transferred payload data. The selected algorithms NNF, NNST, EDCD2, LRMV, and FICA are evaluated using real-time data sets. The performance indicators included are energy consumption, data accuracy, and data reduction percentage.

The organization of the article is as follows: Section I presents the introduction and related work of this study and the main contributions. The selected data reduction algorithms are described in Section II. Section III explores the real-time datasets used in this study. Section IV describes the performance metrics used in this study to evaluate the algorithms. Section V presents the study simulation and results. Lastly, Section VI concludes the outcome of the study.

## II. SELECTED DATA REDUCTION ALGORITHMS

### A. Data Reduction based - Neural Network Fitting (NNF) Algorithm

The NNF model provided by MathWorks [12], helps in solving a data fitting problem using a two-layer feed-forward network. It helps in selecting the data, partitioning it into training, validation, and testing sets, defining the network architecture, and training the network.

In this section, the application of the NNF model to reduce the size of data transferred is described in detail. Fig. 1 represents the block diagram of WSN data reduction based on the NNF algorithm with a general structure. In the training phase, first select the sensor  $S1(t)$  with the highest correlation attribute as the input data of the NNF model and the other sensor features  $S2(t)$  and  $S3(t)$  as the output target of the NNF. The main objective in training NNF is to predict the values  $PS2(t)$  and  $PS3(t)$  from a single input sensor  $S1(t)$  during the reduction phase. As mentioned earlier, NNF is used to decrease the size of the transmitted data by the sensor node.

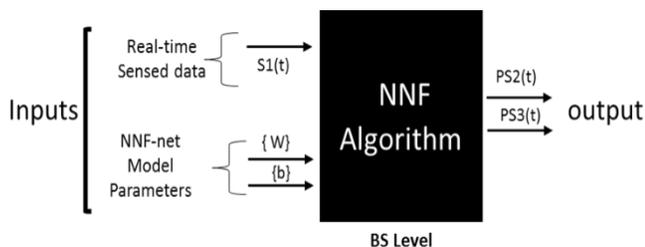


Fig. 1. General Block Diagram of NNF Algorithm.

The detailed description of the data reduction based NNF algorithm is stated by means of the following pseudocode.

#### NNF Model

```

1 Input: Inputs, targets
2 Output: Net // NNF Model with
3 Begin:
4 // Phase I : Create a Fitting Network //
5 Set Hidden Layer Size ← 10
6 Set net ← Call FITNET (Hidden Layer Size)
7 Select InputOutput Pre-Post-Processing-Functions // n=1, m=2
8 Set net.Inputs{n}.process-Fcns ← 'removeconstantrows','mapminmax'
9 Set net.outputs{m}.process-Fcns ← 'removeconstantrows','mapminmax'
10 // Phase II : Setup Division of Data for Training, Validation,
 Testing //
11 Set TrainRatio, ValRatio and TestRatio ← {70,15,15}
12 Assign the training function // "Levenberg-Marquardt backpropagation
13 Select a Performance_Function // ""
14 Set NETPERFORMFCN ← 'mse'; % Mean squared error
15 Phase III // Network -Train
16 Set [net,tr] ← Call TRAIN(net,inputs,targets)
17 End

```

#### NNF algorithm

```

1 Input: S1(t), S2(t), S3(t) // Sensor value for S1 // real-time data
2 Output: PS2(t), PS3(t)
3 Begin:
4 Call NNF Model
5 For i=1 to M do // M is the number of samples
6 Send S1(i) → BS // Send vaule of the sensor 1 To BS
7 // At BS //
8 Estimate [PS2(i), PS3(i)] ← NNF(S1(i)) // Estimated data by
 NNF
9 // Calculate error // This step for check the performance of the
 algorithm
10 Err2(i) ← ABS(PS2(i) – S2(i))
11 Err3(i) ← ABS(PS3(i) – S3(i))
12 End
13 End

```

### B. Data Reduction based - Neural Network Time Series (NNTS) Algorithm

The prediction model NNTS provides by MathWorks [12]. NNTS is a type of dynamic filtering, that uses past-values of one or more-time series to predict future values. Dynamic neural networks containing tapped delay lines are used for nonlinear filtering and prediction.

This section describes in detail the NNTS algorithm used to reduce the amount of data transmitted. Fig. 2 represents the block diagram of WSN data reduction based on the NNTS algorithm with a general structure. In the training phase, first select the sensor  $S1(t)$  with a high correlation attribute as the input data of the NNTS model and the other sensor features  $S2(t)$  and  $S3(t)$  as the output target of NNTS. The main objective in training NNTS is to predict the values  $PS2(t)$  and  $PS3(t)$  from a single input sensor  $S1(t)$  during the reduction phase, where  $S1(t-1)$  and  $S1(t-2)$  are the last two received values of sensor  $S1(t)$ . As mentioned earlier, NNTS is used to decrease the size of the transmitted data by the sensor node.

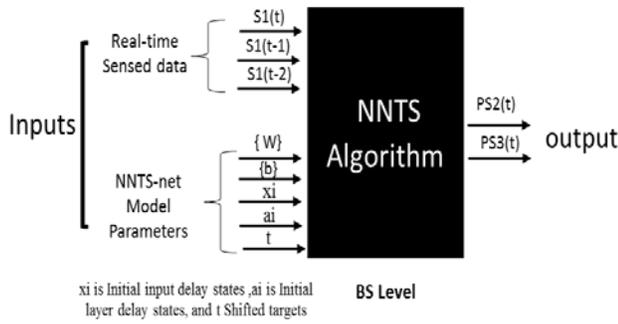


Fig. 2. General Block Diagram of NNTS Algorithm.

The detailed description of the data reduction based NNTS algorithm is given in the following pseudocode.

**// Training NNTS Model//**

```

1 Input: Inputs,Targets
2 Output: Net // NNTS Model
3 Begin:
4 // Phase I : InputOutput TimeSeries Problem with a Time Delay
Neural Network//
5 Convert data_to_ standard NNet cell array form using Tonndata_Fun
6 Set X ← Tonndata (Inputs,false,false)
7 Set T ← Tonndata (Targets, false, false)
8 Create a Time Delay Network
9 Set Input_Delays ← 1:2
10 Set Hidden_LayerSize ← 10
11 Assign the trainingfunction //” Levenberg-Marquardt backpropagation”
12 TrainFcn ← 'trainlm'
13 Set net ← Call Timedelaynet(InputDelays,HiddenLayerSize,TrainFcn)
14 Select Input_Output Pre-Post-ProcessingFunctions//
15 Set net_inputs .process_Fcns←'removeconstantrows','mapminmax'
16 Set net_outputs .Process_Fcns←'removeconstantrows','mapminmax'
17 Prepare TrainingData using Preparat_Fun
18 Set [x,xi,ai,t] ← Prepares(net,X,T) // x is Shifted inputs, xi is Initial
inputdelay states ,ai is Initial_layer_ delay_states, and t Shifted
targets
19 // Phase II: Setup Division of Data for Training, Validation,
Testing//
20 Set Train_Ratio, Val_Ratio and Test_Ratio ← {70, 15, 15}
21 Select a Performanc_ Function//
22 Set NET_PERFORMFCN ← 'mse'; % Mean squared error
23 Phase III // Train the Network
24 Set [net,tr] ← Call TRAINFUN (net,x,t,xi,ai)
25 End

```

**NNTS algorithm**

```

1 Input: S1(t), S2(t), S3(t) // Sensor values // real-time data
2 Output: PS2(t), PS3(t)
3 Begin:
4 Call NNTS Model
5 For i=1 to M do // M is the number of samples
6 Send S1(i) → BS // Send vaule of the sensor 1 To BS
7 // At BS //
8 Set xi ← ([S1(i-1), S1(i-2)])// The recent two received values of the
sensor 1
9 Determine PS2(i), PS3(i) ← NNTS(S1(i), xi,ai) // Estimated data
by NNTS
10 // Calculate error // This step for check the performance of the
algorithm
11 Err2(i) ← ABS(PS2(i) – S2(i))
12 Err3(i) ← ABS(PS3(i) – S3(i))
13 End for
14 End Algorithm

```

**C. Data Reduction based – Linear Regression with Multiple Variables (LRMV) Algorithm**

In statistics, linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables (also referred to as dependent and independent variables). The theoretical concept of using linear regression with multiple variables was explained in detail by Ng, Andrew in [13].

In this section, the application of the LRMV algorithm to reduce the amount of data transferred is described in detail. Fig. 3 represents the block diagram of WSN data reduction based on the LRMV algorithm with a general structure, where the sensor S1(t) is assigned as the dependent variable of the LRMV model, and the other sensor features S2(t) and S3(t) are assigned as the predictor/independent variables of LRMV during the training phase. The aim of training LRMV is to predict the PS1(t) value from multiple sensors S2(t) and S3(t) during the reduction phase. The LRMV parameters are theta (θ), mean (mu), and standard deviation (SSDV). As mentioned earlier, LRMV the size of the transmitted data by the sensor node.

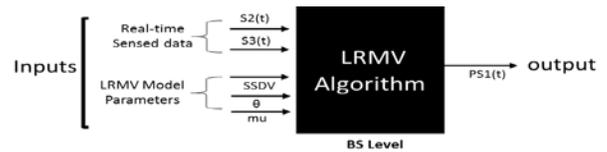


Fig. 3. General Block Diagram of LRMV Algorithm.

The detailed description of the data reduction based LRMV algorithm is stated in the following pseudocode.

```

// Training LRMV Model//
1 Input: Training data // Sensed data S1, S2, S3// Features
2 Output: theta - θ , Mean - μ and standard deviation - SSDV // LRMV
3 Model parameters
4 Begin
5 // Phase I: Load Data and Initialize Variables //
6 Load Sensor dataset // S1, S2, S3
7 Set X \leftarrow [S2 S3]
8 Set Y \leftarrow S1//
9 // Set Initialize Variables //
10 Set M \leftarrow Number of training samples//
11 Set D \leftarrow Number of features//
12 Initialize $\theta \leftarrow$ ZEROS(D+1,1); // Initialize thetas to zero.
13 Initialize Number_itr \leftarrow N; // Set the number of repetitions for gradient
 descent.
14 Initialize $\alpha \leftarrow$ 0.5; // Set alpha Learning rate
15 // Phase II: // Calculate Theta from Normal Equation// data
 processing
16 Set XNormEqn \leftarrow [ones(M,1) X]
17 Calculate thetaNormEqn \leftarrow Call NormalEquation(XNormEqn,Y)
18 //Phase III // Feature Normalization//
19 Normalizing Features for gradient descent
20 Calculate [X, μ , stddev] \leftarrow Call featureNormalize(X)
21 Calculating β via gradient descent
22 $\theta \leftarrow$ Call gradientDescent(X, θ , Y, α , Number_itr)
23 End

```

**LRMV algorithm**

```

1 Input: S1(t), S2(t), S3(t) // Sensor values // real-time data
2 Output: PS1(t)
3 Begin:
4 [θ , μ , stddev] \leftarrow Call LRMV Model
5 For i=1 to M do // M is the number of samples
6 Send S2(i), S3(i) \rightarrow BS // Send vaule of the two sensors 2,3 To BS
7 // At BS //
8 Set X \leftarrow ([S1(i), S2(i)])// The recent received values of the sensor 2 and
 3
9 Determine PS1(i) \leftarrow LRMV_Prediction_Fun (X, θ , μ , stddev) //
 Estimated data by LRMV Prediction Fun
10 // Calculate error // This step for check the performance of the
 algorithm
11 Err1(i) \leftarrow ABS(PS1(i) - S1(i))
12 End for
13 End

```

**D. Data Reduction based –EDCD2 Algorithm**

EDCD2 is a scheme to bring up-to-date measured data to the BS [14]. EDCCD2 was used to decrease the number of transferred packets from nodes (multiple sensors). It should be noted that there are two versions of EDCCD, EDCCD1, and EDCCD2 for sensor nodes with one and multiple sensors, respectively.

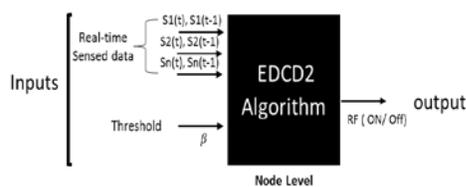


Fig. 4. General Block Diagram of EDCCD2 Algorithm.

In this section, the application of the EDCCD2 to reduce the size of data transferred is described in detail. Fig. 4 shows the block diagram of WSN data reduction based on the EDCCD2 algorithm with a general structure. The basic idea of EDCCD2 is to avoid transmitting the sensed data if the value of the relative difference between the currently sensed data  $S(t)$  and the last transmitted data  $S(t-1)$  is smaller than the threshold value  $\beta$  for all sensors of the same node, otherwise, the sensed data  $S(t)$  will be transmitted to the BS. The detailed description of the EDCCD2 algorithm based on data reduction is given in the following pseudocode.

```

//EDCCD2//
1 Inputs:
 Si(t), Si(t - 1) for each sensor Si and β .
2 Output: Ds
3 Begin:
4 For i = 1: n Do // i=1, 2,...n ;
5 Set Si(t - 1) \leftarrow last measuring value transmitted by the sensor Si
6 Read: the sensor value (SVi) at t time
7 Set Si(t) \leftarrow (SVi)
8 //Calculate the relative differences (Rf)
9 Rf = Abs (S(t) - S(t - 1)) / (S(t) + S(t - 1)) \times 0.5)
10 If Rf > β Then
11 Set SSi \leftarrow 1
12 Else: Set SSi \leftarrow 0
13 End if
14 End For
15 // Recalculate the node data size (Ds)
16 Set Ds \leftarrow 0;
17 For i = 1: n Do
18 Ds = (Ds + (SVi \times SSi))
19 End For
20 // The decision to send data
21 If Ds = 0 Then
22 RFtransmit (Off) // no update / no send
23 Else: RFtransmit (On) // update(send)
24 End If
25 End Algorithm

```

**E. Data Reduction based – Fast Independent Component Analysis (FICA) Algorithm**

Fast Independent Component Analysis (FICA) is an efficient and popular algorithm for independent component analysis developed by Aapo Hyvaerinen at Helsinki University of Technology. [15][16]. Like most FICA algorithms, FICA searches for an orthogonal rotation of the previously whitened data through a fixed-point iteration scheme that make the most of a measure of the non-Gaussian distribution of the rotated components.

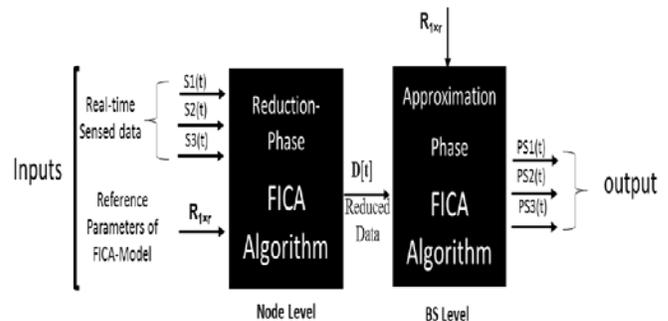


Fig. 5. General Block Diagram of FICA Algorithm.

This section provides a detailed description of the FICA algorithm to reduce the amount of data transferred. Fig. 5 shows the block diagram of WSN data reduction based on the FICA algorithm with a general structure consisting of two phases, namely the reduction phase at the sensor node level and the approximation phase at the BS level. The main objective of training the FICA model is to determine the reference parameters  $R_{(1 \times r)}$ , which are then stored on the sensor node and the same copy is stored on BS. At the node level, the new data  $S1(t)$ ,  $S2(t)$ , and  $S3(t)$  acquired in real-time are reduced by applying the FICA algorithm before transmission and then the reduced data  $D(t)$  is sent to BS. After that, the originally acquired data  $PS1(t)$ ,  $PS2(t)$ , and  $PS3(t)$  are estimated by the approximation phase at BS by applying FICA with the same reference parameters  $R(1 \times r)$  used to reduce the node-level data. As mentioned earlier, FICA is used to reduce the packet size of the sensor node. The detailed description of the data reduction-based FICA algorithm is given in the following pseudocode

```
// FICA
1 Inputs: Training data $\bar{S}_{m \times n}[T] // S1, S2, S3$
2 Output: $D(t)$
3 Begin:
4 Load the training data $\bar{S}_{m \times n}[T]$.
5 Apply FICA ($\bar{S}_{m \times n}[T]$) and calculate the eigenvector array $R_{n \times r}$.
 Then, decreases the eigenvector array to $R_{1 \times r}$, and preserved a copy at the node and transfers one copy to the BS.
6 Standardizes the current measured data $\bar{S}_{1 \times n}[t]$, then decreases it before transferring via the following Equation:

$$D_{1 \times r}[t] = \bar{S}_{1 \times n}[t] \times R_{1 \times r}$$

7 Send the diminished data $D_{1 \times r}[t]$ to the BS.
8 Approximations data at BS by applying

$$\hat{S}_{1 \times n}[t] = D_{1 \times r}[t] \times R_{r \times n}$$

9 End Algorithm
```

### III. REAL-TIME DATASETS

The considered algorithms are evaluated on different benchmark real-time datasets, as described in the following subsections. It's important to note that, usually only part of the data from specific nodes of these datasets are used to assess the performance of current data reduction methods in WSN [17][18][19][20][11][21][22][23]. The reason is that most data reduction methods focus on reducing the amount of transferred data without considering how this data is forwarded to the CH /BS. In other words, they assume that the sensor nodes can directly transmit the sensed data to the CH /BS. The selected algorithms NNF, NNST, EDCD2, LRMV, and FICA are evaluated on real-time datasets as shown below:

#### A. Data I-AirQ

Data I- Air Quality (AirQ) is a WSN data set, including air pressure, humidity and temperature sensors. These sensor data have been collected by 56 sensor nodes in year 2017 at Krakow, Poland. For more information, see the main source [24]. Fig. 6 shows the structure of Data1- AirQ. In addition, some samples of sensors value provided in Tables I to V.

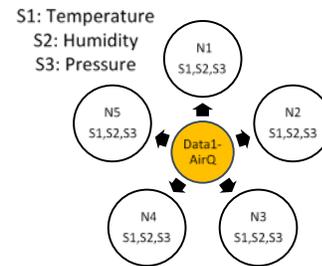


Fig. 6. Structure of Data1- AirQ.

TABLE I. SOME DATA SAMPLES OF NODE1 – DATA1- AIRQ

| Sample | Temperature | Humidity | Pressure |
|--------|-------------|----------|----------|
| 1      | 6           | 92       | 101906   |
| 2      | 6           | 92       | 101869   |
| 3      | 5           | 94       | 101837   |
| 4      | 5           | 92       | 101834   |
| 5      | 4           | 94       | 101832   |
| 6      | 5           | 94       | 101833   |
| 7      | 9           | 78       | 101842   |
| 8      | 11          | 66       | 101831   |
| 9      | 15          | 50       | 101798   |
| 10     | 17          | 42       | 101745   |

TABLE II. SOME DATA SAMPLES OF NODE2 – DATA1- AIRQ

| Sample | Temperature | Humidity | Pressure |
|--------|-------------|----------|----------|
| 1      | 15          | 90       | 101514   |
| 2      | 15          | 90       | 101516   |
| 3      | 15          | 90       | 101530   |
| 4      | 15          | 92       | 101555   |
| 5      | 16          | 90       | 101577   |
| 6      | 16          | 90       | 101595   |
| 7      | 19          | 91       | 101601   |
| 8      | 19          | 88       | 101592   |
| 9      | 20          | 82       | 101571   |
| 10     | 21          | 81       | 101541   |

TABLE III. SOME DATA SAMPLES OF NODE3 – DATA1- AIRQ

| Sample | Temperature | Humidity | Pressure |
|--------|-------------|----------|----------|
| 1      | 14          | 88       | 100825   |
| 2      | 14          | 92       | 100797   |
| 3      | 14          | 94       | 100781   |
| 4      | 14          | 94       | 100805   |
| 5      | 14          | 92       | 100761   |
| 6      | 14          | 90       | 100795   |
| 7      | 15          | 92       | 100822   |
| 8      | 15          | 90       | 100839   |
| 9      | 16          | 85       | 100834   |
| 10     | 18          | 77       | 100849   |

TABLE IV. SOME DATA SAMPLES OF NODE4 – DATA1- AIRQ

| Sample | Temperature | Humidity | Pressure |
|--------|-------------|----------|----------|
| 1      | 8           | 104      | 101967   |
| 2      | 7           | 109      | 101969   |
| 3      | 6           | 112      | 101975   |
| 4      | 6           | 114      | 101980   |
| 5      | 6           | 112      | 101963   |
| 6      | 8           | 105      | 101920   |
| 7      | 8           | 99       | 101895   |
| 8      | 10          | 87       | 101864   |
| 9      | 11          | 82       | 101837   |
| 10     | 12          | 78       | 101853   |

TABLE V. SOME DATA SAMPLES OF NODE5 – DATA1- AIRQ

| Sample | Temperature | Humidity | Pressure |
|--------|-------------|----------|----------|
| 1      | 5           | 94       | 102384   |
| 2      | 4           | 100      | 102396   |
| 3      | 3           | 94       | 102413   |
| 4      | 3           | 100      | 102415   |
| 5      | 2           | 97       | 102453   |
| 6      | 2           | 94       | 102510   |
| 7      | 4           | 94       | 102564   |
| 8      | 6           | 88       | 102593   |
| 9      | 9           | 82       | 102606   |
| 10     | 11          | 76       | 102603   |

B. Data 2-ARHO

Data2- American River Hydrologic Observatory (ARHO) is a WSNs data set, including soil temperature, relative humidity, snow depth sensors, etc. These sensor data have been collected by 130 spatially distributed sensor nodes in the river basin of the United States. Period: the water year 2014 to the water year 2017. For more information, see the main source [25]. Fig. 7 shows the structure of Data 2- ARHO, also some examples of sensor values for all used nodes in Tables VI to X.

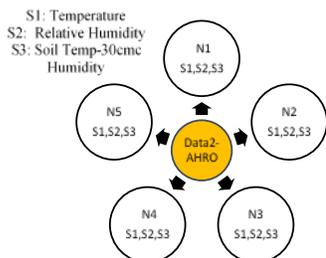


Fig. 7. Structure of Data-2 ARHO.

TABLE VI. SOME DATA SAMPLES OF NODE1 – DATA2- ARHO

| Sample | Temperature (°C) | Relative Humidity (%) | Soil Temp-30cmc |
|--------|------------------|-----------------------|-----------------|
| 1      | 13.2             | 38.462                | 11.5            |
| 2      | 12.52            | 39.867                | 11.6            |
| 3      | 12.01            | 40.756                | 11.6            |
| 4      | 11.45            | 42.309                | 11.7            |
| 5      | 11.04            | 43.095                | 11.8            |
| 6      | 10.16            | 45.276                | 11.9            |
| 7      | 9.52             | 46.607                | 11.9            |
| 8      | 8.98             | 47.843                | 12              |
| 9      | 8.31             | 47.911                | 12              |
| 10     | 7.87             | 50.451                | 12              |

TABLE VII. SOME DATA SAMPLES OF NODE2 – DATA2- ARHO

| Sample | Temperature (°C) | Relative Humidity (%) | Soil Temp-30cmc |
|--------|------------------|-----------------------|-----------------|
| 1      | 13.91            | 37.442                | 15.3            |
| 2      | 13.65            | 38.112                | 15.6            |
| 3      | 13.18            | 39.249                | 15.9            |
| 4      | 12.54            | 41.042                | 16.1            |
| 5      | 12.2             | 42.072                | 16.4            |
| 6      | 11.16            | 44.384                | 16.6            |
| 7      | 10.39            | 46.591                | 16.8            |
| 8      | 9.38             | 49.097                | 16.9            |
| 9      | 8.81             | 48.684                | 17.1            |
| 10     | 8.4              | 50.978                | 17.2            |

TABLE VIII. SOME DATA SAMPLES OF NODE3 – DATA2- ARHO

| Sample | Temperature (°C) | Relative Humidity (%) | Soil Temp-30cmc |
|--------|------------------|-----------------------|-----------------|
| 1      | 13.44            | 40.587                | 13.9            |
| 2      | 12.76            | 42.232                | 14.1            |
| 3      | 12.07            | 44.779                | 14.1            |
| 4      | 11.66            | 46.026                | 14.2            |
| 5      | 11.24            | 47.229                | 14.4            |
| 6      | 10.72            | 48.717                | 14.5            |
| 7      | 10.16            | 49.661                | 14.5            |
| 8      | 9.9              | 50.147                | 14.6            |
| 9      | 9.43             | 50.694                | 14.7            |
| 10     | 8.86             | 51.858                | 14.7            |

TABLE IX. SOME DATA SAMPLES OF NODE4 – DATA2- ARHO

| Sample | Temperature (°C) | Relative Humidity (%) | Soil Temp-30cmc |
|--------|------------------|-----------------------|-----------------|
| 1      | 13.46            | 38.555                | 11.1            |
| 2      | 13.12            | 39.243                | 11.1            |
| 3      | 12.56            | 40.556                | 11.1            |
| 4      | 12.07            | 41.831                | 11.2            |
| 5      | 11.66            | 43.487                | 11.2            |
| 6      | 11.34            | 44.469                | 11.2            |
| 7      | 10.92            | 45.431                | 11.3            |
| 8      | 10.37            | 46.87                 | 11.3            |
| 9      | 9.84             | 48.325                | 11.3            |
| 10     | 9.46             | 47.626                | 11.3            |

TABLE X. SOME DATA SAMPLES OF NODE5 – DATA2- ARHO

| Sample | Temperature (°C) | Relative Humidity (%) | Soil Temp-30cmc |
|--------|------------------|-----------------------|-----------------|
| 1      | 13.12            | 39.374                | 11.4            |
| 2      | 12.58            | 40.493                | 11.4            |
| 3      | 12.33            | 41.796                | 11.5            |
| 4      | 11.98            | 42.819                | 11.5            |
| 5      | 11.04            | 44.845                | 11.6            |
| 6      | 10.66            | 45.966                | 11.6            |
| 7      | 10.3             | 46.642                | 11.6            |
| 8      | 9.64             | 46.653                | 11.7            |
| 9      | 9.25             | 48.156                | 11.7            |
| 10     | 8.95             | 49.285                | 11.7            |

C. Data 3- GSB

Data3- “Grand St. Bernard (GSB)” is WSN data set, which it was gathered by deployed 23 sensors to observe the measurement characteristics of the environmental in the “Grand Saint Bernard Pass” between Switzerland and Italy. The sensors are relative humidity, surface temperature, and ambient temperature [26]. Fig. 8 shows the structure of Data 3- GSB. Some examples of the sensor values of all the nodes used can be found in Tables XI to XV.

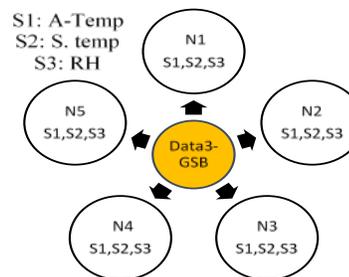


Fig. 8. Structure of Data-3 GSB.

TABLE XI. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE1\_ DATA 3- GSB

| Sample | A. Temperature (°C) | Relative Humidity (%) | S. Temperature (°C) |
|--------|---------------------|-----------------------|---------------------|
| 1      | 12.915              | 51.7785               | 12.881              |
| 2      | 12.53               | 52.3073               | 12.183              |
| 3      | 12.56               | 50.2515               | 12.5995             |
| 4      | 13.1533             | 51.1487               | 13.8287             |
| 5      | 12.65               | 51.121                | 13.131              |
| 6      | 12.81               | 51.4583               | 13.7457             |
| 7      | 12.52               | 51.2055               | 12.412              |
| 8      | 12.6267             | 50.2657               | 12.475              |
| 9      | 12.52               | 50.19                 | 12.412              |
| 10     | 12.59               | 51.2353               | 12.058              |

TABLE XII. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE2\_ DATA 3- GSB

| Sample | A. Temperature (°C) | Relative Humidity (%) | S. Temperature (°C) |
|--------|---------------------|-----------------------|---------------------|
| 1      | 6.48                | 84.963                | 4.225               |
| 2      | 6.36                | 85.505                | 4.537               |
| 3      | 6.3                 | 86.08                 | 5.35                |
| 4      | 6.23                | 86.558                | 5.975               |
| 5      | 6.18                | 86.904                | 6.475               |
| 6      | 6.18                | 87.128                | 6.6                 |
| 7      | 6.16                | 87.257                | 6.725               |
| 8      | 6.16                | 87.39                 | 6.787               |
| 9      | 6.15                | 87.499                | 6.85                |
| 10     | 6.22                | 87.649                | 6.787               |

TABLE XIII. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE3\_ DATA 3- GSB

| Sample | A. Temperature (°C) | Relative Humidity (%) | S. Temperature (°C) |
|--------|---------------------|-----------------------|---------------------|
| 1      | 10.92               | 87.232                | 11.412              |
| 2      | 10.94               | 87.877                | 11.35               |
| 3      | 10.9                | 87.433                | 11.412              |
| 4      | 10.9                | 87.296                | 11.35               |
| 5      | 10.88               | 86.946                | 11.287              |
| 6      | 10.91               | 87.436                | 11.225              |
| 7      | 10.93               | 88.057                | 11.225              |
| 8      | 10.91               | 88.211                | 11.287              |
| 9      | 10.85               | 87.192                | 11.162              |
| 10     | 10.88               | 87.406                | 11.35               |

TABLE XIV. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE4\_ DATA 3- GSB

| Sample | A. Temperature (°C) | Relative Humidity (%) | S. Temperature (°C) |
|--------|---------------------|-----------------------|---------------------|
| 1      | 0.474802            | 7.08262               | 0.274989            |
| 2      | 0.482715            | 7.076051              | 0.33434             |
| 3      | 0.473219            | 7.095914              | 0.413473            |
| 4      | 0.471637            | 7.146639              | 0.47773             |
| 5      | 0.477176            | 7.171882              | 0.507405            |
| 6      | 0.474011            | 7.172753              | 0.53708             |
| 7      | 0.476385            | 7.168242              | 0.561849            |
| 8      | 0.481133            | 7.193169              | 0.561849            |
| 9      | 0.474011            | 7.231628              | 0.561849            |
| 10     | 0.471637            | 7.247771              | 0.561849            |

TABLE XV. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE5\_ DATA3- GSB

| Sample | A. Temperature (°C) | Relative Humidity (%) | S. Temperature (°C) |
|--------|---------------------|-----------------------|---------------------|
| 1      | 4.66891             | 37.29656              | 4.879268            |
| 2      | 4.677461            | 37.57233              | 4.852759            |
| 3      | 4.660359            | 37.38249              | 4.879268            |
| 4      | 4.660359            | 37.32392              | 4.852759            |
| 5      | 4.651808            | 37.17427              | 4.825823            |
| 6      | 4.664635            | 37.38378              | 4.799315            |
| 7      | 4.673186            | 37.64929              | 4.799315            |
| 8      | 4.664635            | 37.71513              | 4.825823            |
| 9      | 4.638981            | 37.27945              | 4.772379            |
| 10     | 4.651808            | 37.37095              | 4.852759            |

D. Data 4- Intel

Data4- “Intel Berkeley Research Lab (IBRL)” is a WSN data-set, which it was gathered by deployed 54 Mica2Dot sensor nodes at “Intel’s research Lab”, University of Berkeley. The wireless network consisted of. The WSN includes various sensors: voltage, temperature, light, and humidity [27]. Fig. 9 shows the structure of Data 4- Intel, also some examples of sensor values for all the nodes used in Tables VI to XX.

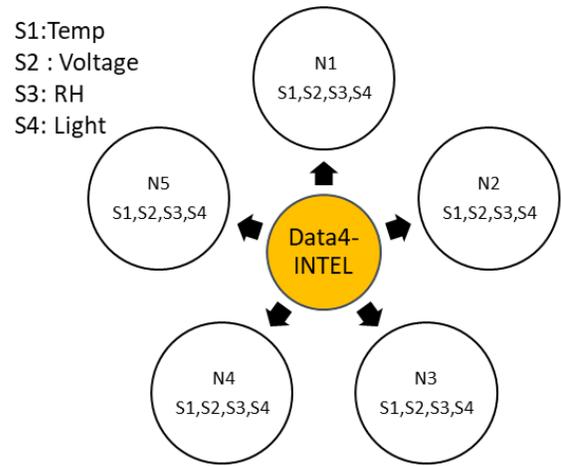


Fig. 9. Structure of Data4-Intel.

TABLE XVI. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE1\_ DATA4-INTEL

| Sample | Temperature (°C) | Relative Humidity (%) | Light (Lux) | Voltage (V) |
|--------|------------------|-----------------------|-------------|-------------|
| 1      | 19.9884          | 37.0933               | 45.08       | 2.034       |
| 2      | 19.9884          | 37.0933               | 45.08       | 2.6996      |
| 3      | 19.3024          | 38.4629               | 45.08       | 2.6874      |
| 4      | 19.1652          | 38.8039               | 45.08       | 2.6874      |
| 5      | 19.175           | 38.8379               | 45.08       | 2.6996      |
| 6      | 19.1456          | 38.9401               | 45.08       | 2.6874      |
| 7      | 19.1652          | 38.872                | 45.08       | 2.6874      |
| 8      | 19.1652          | 38.8039               | 45.08       | 2.6874      |
| 9      | 19.1456          | 38.8379               | 45.08       | 2.6996      |
| 10     | 19.1456          | 38.872                | 45.08       | 2.6874      |

TABLE XVII. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE2 – DATA4-INTEL

| Sample | Temperature (°C) | Relative Humidity (%) | Light (Lux) | Voltage (V) |
|--------|------------------|-----------------------|-------------|-------------|
| 1      | 89.5488          | 29.2581               | 11.96       | 1.9537      |
| 2      | 19.567           | 39.6878               | 121.44      | 2.6753      |
| 3      | 19.5376          | 39.7557               | 121.44      | 2.6753      |
| 4      | 19.4788          | 39.6878               | 121.44      | 2.6633      |
| 5      | 19.4494          | 39.7217               | 121.44      | 2.6753      |
| 6      | 19.4984          | 39.586                | 121.44      | 2.6753      |
| 7      | 19.4788          | 39.5521               | 121.44      | 2.6633      |
| 8      | 19.4592          | 39.5181               | 121.44      | 2.6753      |
| 9      | 19.4494          | 39.5521               | 121.44      | 2.6753      |
| 10     | 19.4788          | 39.4162               | 121.44      | 2.6874      |

TABLE XVIII. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE3 – DATA4-INTEL

| Sample | Temperature (°C) | Relative Humidity (%) | Light (Lux) | Voltage (V) |
|--------|------------------|-----------------------|-------------|-------------|
| 1      | 22.2816          | 43.8515               | 41.4        | 2.5935      |
| 2      | 22.2718          | 43.9844               | 41.4        | 2.5935      |
| 3      | 22.2718          | 43.8848               | 41.4        | 2.5935      |
| 4      | 22.5168          | 48.1243               | 382.72      | 2.32        |
| 5      | 22.2816          | 43.918                | 41.4        | 2.5935      |
| 6      | 22.2718          | 43.7186               | 41.4        | 2.5935      |
| 7      | 22.262           | 43.519                | 41.4        | 2.6049      |
| 8      | 22.2718          | 43.519                | 41.4        | 2.5935      |
| 9      | 22.2718          | 43.7186               | 41.4        | 2.5935      |
| 10     | 22.6148          | 47.7013               | 264.96      | 2.32        |

TABLE XIX. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE4 – DATA4-INTEL

| Sample | Temperature (°C) | Relative Humidity (%) | Light (Lux) | Voltage (V) |
|--------|------------------|-----------------------|-------------|-------------|
| 1      | 19.1848          | 38.9742               | 108.56      | 2.6874      |
| 2      | 19.0084          | 39.4502               | 108.56      | 2.6874      |
| 3      | 18.9398          | 39.6539               | 108.56      | 2.6996      |
| 4      | 18.8712          | 40.0607               | 108.56      | 2.6874      |
| 5      | 18.8516          | 40.0945               | 108.56      | 2.6996      |
| 6      | 18.8418          | 40.2976               | 108.56      | 2.6996      |
| 7      | 18.8418          | 40.2976               | 108.56      | 2.6996      |
| 8      | 18.832           | 40.2976               | 108.56      | 2.6996      |
| 9      | 18.8222          | 40.2638               | 108.56      | 2.6996      |
| 10     | 18.8124          | 40.2976               | 108.56      | 2.6874      |

TABLE XX. SOME TEMPERATURE SENSOR DATA SAMPLES OF NODE5 – DATA4-INTEL

| Sample | Temperature (°C) | Relative Humidity (%) | Light (Lux) | Voltage (V) |
|--------|------------------|-----------------------|-------------|-------------|
| 1      | 19.3612          | 39.8235               | 75.44       | 2.67532     |
| 2      | 19.273           | 39.9252               | 75.44       | 2.67532     |
| 3      | 18.9888          | 40.9392               | 75.44       | 2.67532     |
| 4      | 18.9398          | 40.8718               | 75.44       | 2.67532     |
| 5      | 18.9006          | 40.973                | 75.44       | 2.68742     |
| 6      | 18.9496          | 40.9055               | 75.44       | 2.67532     |
| 7      | 18.9594          | 41.3098               | 75.44       | 2.67532     |
| 8      | 18.9496          | 41.3771               | 75.44       | 2.67532     |
| 9      | 18.9496          | 41.0404               | 75.44       | 2.67532     |
| 10     | 18.93            | 40.9055               | 75.44       | 2.67532     |

#### IV. PERFORMANCE METRICS

##### A. Accuracy

Accuracy is the overall average of absolute error for all selected nodes from the same dataset as defined below:

$$Accuracy = \frac{\sum_{k=1}^L AEPN(k)}{L} \quad (1)$$

$$AEPN(k) = \frac{\sum_{i=1}^N |AEPS(i)|}{N}, \quad (2)$$

$$AEPS(i) = \frac{\sum_{j=1}^M |SV(j) - RV(j)|}{M} \quad (3)$$

Where  $K = \{1, 2, \dots, L\}$ ,  $L$  is the number of nodes.,  $AEPN$  is the average of absolute error for all samples transmitted by the node ( $k$ ),  $SV$  is the sensor value at the sensor node,  $RV$  is the received value at  $BS$ , and  $AEPS(i)$  is the mean Absolute error for sensor ( $i$ ),  $i = \{1, 2, \dots, N\}$ ,  $N$  is the number of sensors,  $M$  is the number of samples transmitted by the node ( $k$ ).

##### B. Data Reduction Ratio %

$$DR\% = \left(1 - \frac{Sd_{redc}}{Sd_{Lenght}}\right) \times 100 \quad (4)$$

where  $DR$  is the ratio of the reduced data,  $Sd_{redc}$  is the size of transferred data after reduction and  $Sd_{Lenght}$  is the size of the unreduced transmission samples.

##### C. Total Energy Consumption

$$TE_{Directly} = D_s \times N_{of\ S} \times C_E PByte \quad (5)$$

$$TE_{DR} = RD_S \times N_{of\ S} \times C_E PByte \quad (6)$$

Where:  $TE_{Directly}$  is the Total Energy consumed in case of the Direct transmission,  $D_s$  is the mean Data size,  $N_{of\ S}$  is the mean Number of Samples,  $C_E PByte$  is the mean Cost Energy Per Byte,  $RD_S$  is the mean Data Reduction.

#### V. SIMULATION AND RESULTS

Fig. 10 shows the accuracy of the applied algorithms ED CD2, FICA, NNF, NNTS, and LRMV for all selected nodes N1, N2, N5 from the DATA1-AIRQ dataset. From the results, the ED CD2 algorithm has the best accuracy compared to the other algorithms FICA, NNF, NNTS, and LRMV. The reason for this is the average total absolute error which has the lowest value of 5.48 when ED CD2 is used for all nodes. Moreover, the algorithms FICA and LRMV have the worst performance in terms of accuracy, and the average absolute errors are 62.30 and 20.13, respectively. Table A1, Table A2, Table A3, Table A5 (see Appendix) showed the average of absolute error for all samples transmitted by the nodes (N1-N5) of the applied algorithms ED CD2, FICA, NNF, NNTS, and LRMV, respectively.

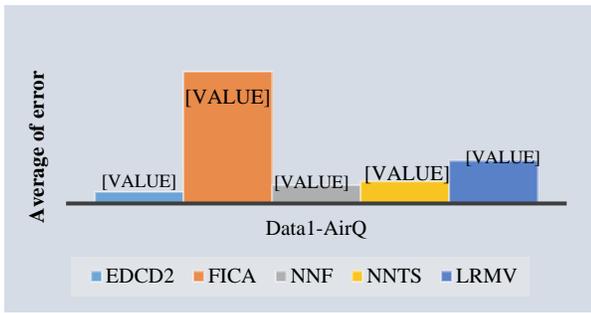


Fig. 10. Accuracy of Applying Various Algorithms for all Selected Nodes from DATA1-AIRQ.

Fig. 11 shows the accuracy of the applied algorithms EDCD2, FICA, NNF, NNTS, and LRMV for all selected nodes N1, N2, N5 from the DATA2-ARHO dataset. From the results, the EDCD2 algorithm has the best accuracy compared to the other algorithms FICA, NNF, NNTS, and LRMV. The reason for this is the average total absolute error which has the lowest value of 0.199 when EDCD2 is used for all nodes. Moreover, NNTS and NNF algorithms have the worst performance in terms of accuracy, and the average absolute errors are 5.38 and 5.62, respectively. In summary, EDCD2 is a threshold-based data reduction algorithm. EDCD2 transmits measurement data only when the relative difference between the current measurement data and the last transmitted data is larger than the threshold value.

Fig. 12 shows the accuracy of the applied EDCD2, FICA, NNF, NNTS, and LRMV algorithms for all selected nodes N1, N2, N5 from the DATA3-GSB dataset. From the results, the EDCD2 algorithm has been shown to have the best accuracy compared with the other algorithms, FICA, NNF, NNTS, and LRMV. The reason is related to the overall average absolute error, which is the lowest value of 0.30 for applied EDCD2 for all nodes. Furthermore, the FICA and NNF algorithms have the worst performance in terms of accuracy, and the average absolute errors are 3.84 and 1.34, respectively.

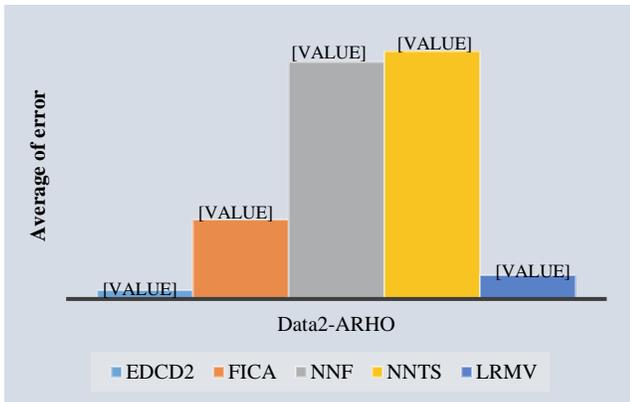


Fig. 11. Accuracy of Applying Various Algorithms for all Selected Nodes from DATA2-ARHO.

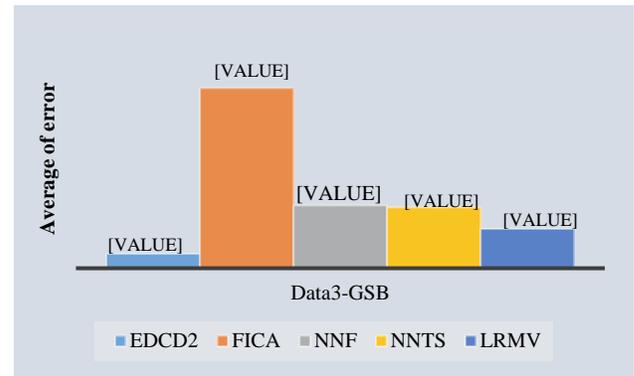


Fig. 12. Accuracy of Applying Various Algorithms for all Selected Nodes from DATA3-GSB.

Fig. 13 shows the accuracy of the applied algorithms EDCD2, FICA, NNF, NNTS, and LRMV for all selected nodes N1, N2, N5 from the DATA4-INTEL dataset. From the results, the NNF algorithm has the best accuracy compared to the other algorithms EDCD2, FICA, NNTS and LRMV. The reason for this is the average total absolute error which has the lowest value of 1.01 when NNF is used for all nodes. Moreover, the algorithms FICA and LRMV have the worst performance in terms of accuracy, and the average absolute errors are 28.68 and 1.54, respectively.

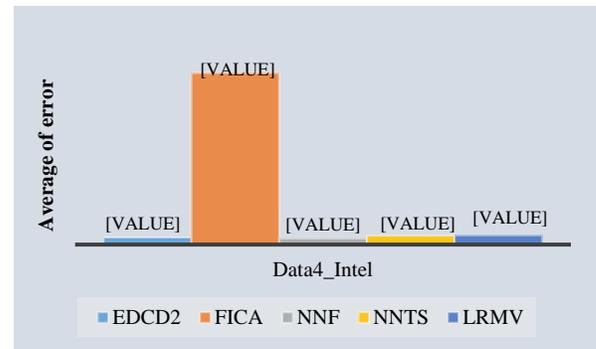


Fig. 13. Accuracy of Applying Various Algorithms for all Selected Nodes from DATA4-INTEL.

Fig. 14 shows the average of data reduction ratio percentage for applying various algorithms for different datasets. The studied algorithms are EDCD2, FICA, NNF, NNTS, and LRMV. The selected datasets are Data1-AirQ, Data2-ARHO, Data3-GSB, and Data4\_Intel. From these results, the average data reduction percentage for applied EDCD2, FICA, NNF, NNTS, and LRMV algorithms through a real-time dataset named Data1-AirQ is 33%, 33%, 67%, 67%, and 33%, respectively. It is noted that the NNF and NNTS algorithms have the highest data reduction. By referring to Fig. 10 both algorithms, NNF and NNTS, have acceptable accuracy and the lowest error has been shown by applying EDCD2. In the same way, the average data reduction percentage for applied NNF, NNTS, EDCD2, LRMV, and FICA algorithms through a real-time dataset named Data2-ARHO is 67%, 67%, 56%, 33%, and 67%, respectively. Although NNF and NNTS algorithms achieved the highest data reduction ratio, both NNF and NNTS have the highest error and worst performance in terms of accuracy as shown in Fig. 11 The average data

reduction percentage for applied NNF, NNTS, EDCD2, LRMV, and FICA algorithms through a real-time dataset named Data3-GSB is 67%, 67%, 67%, 33%, and 33%, respectively. It is noted that the NNF, NNTS and EDCD2 algorithms have the highest data reduction, by referring to Fig. 12 the lowest error has been shown by applying EDCD2. FICA showed the worst performance in terms of accuracy, with the highest errors. The average data reduction percentage

for applied NNF, NNTS, EDCD2, LRMV, and FICA algorithms through a real-time dataset named Data4\_Intel is 50%, 50%, 83%, 25%, and 75%, respectively. It is worth noting that the EDCD2 algorithm achieves the highest data reduction. By referring to Fig. 14, Tables XXI, XXII both NNF, NNTS, EDCD2, and LRMV algorithms have acceptable accuracy, and the highest error has been shown by applying FICA.

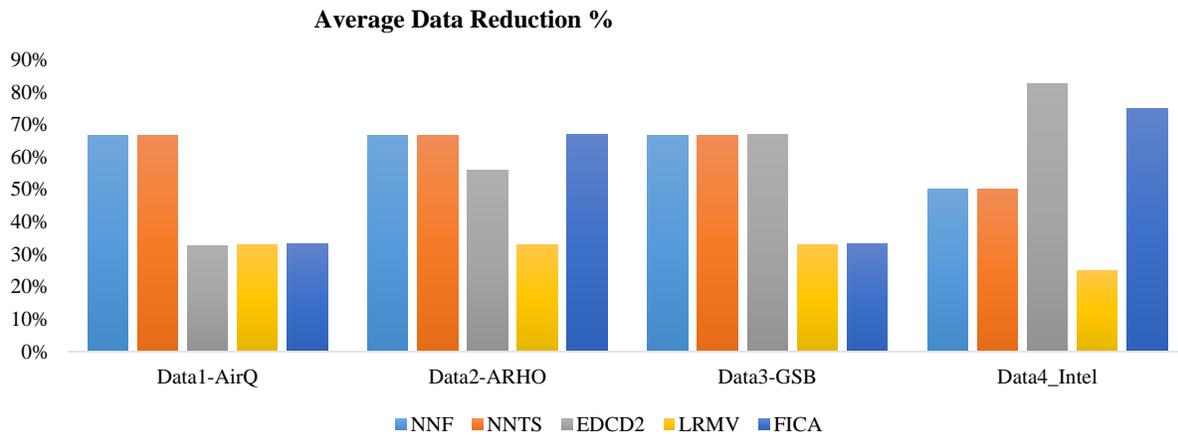


Fig. 14. Average of Data Reduction Ratio % for Applying Various Algorithms for Different Datasets.

TABLE XXI. TOTAL ENERGY CONSUMPTION BY APPLYING THE SELECTED ALGORITHMS VS WITHOUT ALGORITHMS

|             | NNF     | NNTS    | EDCD2    | LRMV    | FICA    | Without algorithm |
|-------------|---------|---------|----------|---------|---------|-------------------|
| Data1-AirQ  | 165900  | 165900  | 335644.2 | 333459  | 331800  | 497700            |
| Data2-ARHO  | 165900  | 165900  | 219718   | 333459  | 164241  | 497700            |
| Data3-GSB   | 711000  | 711000  | 706592.2 | 1429110 | 1422000 | 2133000           |
| Data4_Intel | 1422000 | 1422000 | 491443.2 | 2133000 | 711000  | 2844000           |

TABLE XXII. THE PERCENTAGE OF SAVED ENERGY BY APPLYING THE SELECTED ALGORITHMS

|             | NNF | NNTS | EDCD2 | LRMV | FICA |
|-------------|-----|------|-------|------|------|
| Data1-AirQ  | 67% | 67%  | 33%   | 33%  | 33%  |
| Data2-ARHO  | 67% | 67%  | 56%   | 33%  | 67%  |
| Data3-GSB   | 67% | 67%  | 67%   | 33%  | 33%  |
| Data4_Intel | 50% | 50%  | 83%   | 25%  | 75%  |

## VI. CONCLUSION

The impact of data reduction methods on WSN performance is investigated in this paper, using a set of real-time datasets. Simulation tests are performed in MATLAB for different methods to reduce the amount of data sent. The selected algorithms NNF, NNST, EDCD2, LRMV, and FICA are evaluated using real-time data sets. The performance metrics measured are energy consumption, data accuracy, and percentage of data reduction. The results of the study show that the selected algorithm helps to reduce the amount of transmitted data and energy consumption, and each algorithm performs differently depending on the dataset used.

## ACKNOWLEDGMENT

“The writers would like to thank University Polyethnic of Bucharest (UPB) for their support to carry out this study.”

## REFERENCES

- [1] M. K. Hussein, “Data Reduction Algorithms for Wireless Sensor Networks Applications: Review,” in 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), 2021, pp. 1–7.
- [2] M. I. Husni, M. K. Hussein, N. A. M. Alduais, J. Abdullah, and I. Marghescu, “Performance of Various Algorithms to Reduce the Number of Transmitted Packets by Sensor Nodes in Wireless Sensor Network,” in 2019 11th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 2019, pp. 1–7.
- [3] N. Alduais et al., “An Efficient IoT-based Smart Water Meter System of Smart City Environment,” *Artic. Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, p. 2021, 2021.
- [4] N. A. M. Alduais, J. Abdullah, and A. Jamil, “APRS: adaptive real-time payload data reduction scheme for IoT/WSN sensor board with multivariate sensors,” *Int. J. Sens. Networks*, vol. 28, no. 4, pp. 211–229, 2018.
- [5] A. Jarwan, A. Sabbah, and M. Ibnkahla, “Data Transmission Reduction Schemes in WSNs for Efficient IoT Systems,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1307–1324, 2019.
- [6] S. Diwakaran, B. Perumal, and K. Vimala Devi, “A cluster prediction model-based data collection for energy efficient wireless sensor network,” *J. Supercomput.*, vol. 75, no. 6, pp. 3302–3316, 2019.

[7] E. P. K. Gilbert, B. Kaliaperumal, E. B. Rajsingh, and M. Lydia, "Trust based data prediction, aggregation and reconstruction using compressed sensing for clustered wireless sensor networks," *Comput. Electr. Eng.*, vol. 72, pp. 894–909, 2018.

[8] Y. Fathy, P. Barnaghi, and R. Tafazolli, "An adaptive method for data reduction in the Internet of Things," *IEEE World Forum Internet Things, WF-IoT 2018 - Proc.*, vol. 2018-Janua, pp. 729–735, 2018.

[9] J. Abdullah, M. K. Hussien, N. A. M. Alduais, M. I. Husni, and A. Jamil, "Data Reduction Algorithms based on Computational Intelligence for Wireless Sensor Networks Applications," in *2019 IEEE 9th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 2019, pp. 166–171.

[10] N. A. M. Alduais, J. Abdullah, A. Jamil, and H. Heidari, "Performance Evaluation of Real-Time Multivariate Data Reduction Models for Adaptive-Threshold in Wireless Sensor Networks," *IEEE Sensors Lett.*, vol. 1, no. 6, pp. 1–4, 2017.

[11] N. A. M. ALDUAIS and A. A. J. JIWA ABDULLAH, "RDCM: An Efficient Real-Time Data Collection Model for IoT / WSN Edge With Multivariate Sensors," *IEEE Access*, vol. 7, pp. 89063–89082, 2019.

[12] Fit Data with a Shallow Neural Network - MATLAB & Simulink - MathWorks United Kingdom, "No Title," *Mathworks.com*, 2017. [Online]. Available: <https://www.mathworks.com/help/nnet/gs/fit-data-with-a-neura>, 2017. .

[13] "Linear Regression with Multiple Variables | Machine Learning, Deep Learning, and Computer Vision." [Online]. Available: <https://www.ritchieng.com/multi-variable-linear-regression/>. [Accessed: 25-Sep-2021].

[14] N. A. M. Alduais, "An Efficient Data Collection Algorithms for IoT Sensor Board," 2016.

[15] M. I. Al-Qinna and S. M. Jaber, "Predicting soil bulk density using advanced pedotransfer functions in an arid environment," *Trans. ASABE*, vol. 56, no. 3, pp. 963–976, 2013.

[16] A. Hyvärinen and E. Oja, "A Fast Fixed-Point Algorithm for Independent Component Analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, 1997.

[17] AISSMS Institute of Information Technology and Institute of Electrical and Electronics Engineers, "2020 International Conference on Emerging Smart Computing and Informatics (ESCI): AISSMS Institute of Information Technology, Pune, India. Mar 12-14, 2020.," pp. 103–108, 2020.

[18] L. Mesin, S. Aram, and E. Pasero, "A Neural Data-Driven Approach to increase Wireless Sensor Network(s' lifetime)," pp. 1–3, 2014.

[19] M. A. Rassam, A. Zainal, and M. A. Maarof, "An adaptive and efficient dimension reduction model for multivariate wireless sensor networks applications," *Appl. Soft Comput. J.*, vol. 13, no. 4, pp. 1978–1996, 2013.

[20] H. Harb et al., "Industrial Process Monitoring To cite this version :," 2019.

[21] L. Tan and M. Wu, "Data Reduction in Wireless Sensor Networks: A Hierarchical LMS Prediction Approach," *IEEE Sens. J.*, vol. 16, no. 6, pp. 1708–1715, 2016.

[22] C. Carvalho, D. G. Gomes, N. Agoulmine, and J. N. de Souza, "Improving prediction accuracy for WSN data reduction by applying multivariate spatio-temporal correlation," *Sensors*, vol. 11, no. 11, pp. 10010–10037, 2011.

[23] Y. Deng, C. Han, J. Guo, and L. Sun, "Temporal and spatial nearest neighbor values based missing data imputation in wireless sensor networks," *Sensors*, vol. 21, no. 5, pp. 1–24, 2021.

[24] UCI Machine Learning Repository, "No Title," *Air Quality Data Set*, *Archive.ics.uci.edu*, 2018. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Air+quality>. [Accessed: 23- Jan-2018]., 2018. .

[25] "Dryad Data -- Snow depth, air temperature, humidity, soil moisture and temperature, and solar radiation data from the basin-scale wireless-sensor network in American River Hydrologic Observatory (ARHO)." [Online]. Available: <https://datadryad.org/stash/dataset/doi:10.6071/M39Q2V>. [Accessed: 24-Sep-2021].

[26] "Index of /~rossi/papers/CS\_2012." [Online]. Available: [http://www.dei.unipd.it/~rossi/papers/CS\\_2012/](http://www.dei.unipd.it/~rossi/papers/CS_2012/). [Accessed: 24-Sep-2021].

[27] Intel Lab data, "Intel-dataSet," [Online]. Available: <http://db.csail.mit.edu/labdata/labdata.html>. [Accessed: 23- Jan- 2018]., 2018.

APPENDIX

TABLE A1. AVERAGE OF ABSOLUTE ERROR RESULTS OF APPLYING EDCD2 ALGORITHM FOR ALL NODES WITH DIFFERENT DATASETS

| EDCD2          |              |              |               |              |                  |              |                       |              |                  |                  |                       |              |                  |              |                       |              |              |
|----------------|--------------|--------------|---------------|--------------|------------------|--------------|-----------------------|--------------|------------------|------------------|-----------------------|--------------|------------------|--------------|-----------------------|--------------|--------------|
| Nodes          | Data1-AirQ   |              |               |              | Data2-ARHO       |              |                       |              | Data3-GSB        |                  |                       |              | Data4_Intel      |              |                       |              |              |
|                | Temperature  | Humidity     | Pressure      | Mean         | Temperature (°C) | soilT_30cm   | Relative Humidity (%) | Mean         | Temperature (°C) | Temperature (°C) | Relative Humidity (%) | Mean         | Temperature (°C) | Voltage (v)  | Relative Humidity (%) | Light (Lux)  | Mean         |
| N1             | 0.011        | 0.469        | 20.657        | 7.046        | 0.126            | 0.036        | 0.425                 | 0.196        | 0.093            | 0.112            | 0.565                 | 0.257        | 0.304            | 0.005        | 0.409                 | 1.834        | 0.749        |
| N2             | 0.011        | 0.469        | 20.657        | 7.046        | 0.111            | 0.062        | 0.397                 | 0.190        | 0.092            | 0.113            | 0.823                 | 0.342        | 0.109            | 0.003        | 0.152                 | 1.797        | 0.651        |
| N3             | 0.059        | 0.313        | 8.548         | 2.974        | 0.110            | 0.038        | 0.411                 | 0.186        | 0.100            | 0.107            | 0.780                 | 0.329        | 0.389            | 0.002        | 0.149                 | 0.879        | 0.343        |
| N4             | 0.068        | 0.326        | 10.167        | 3.520        | 0.130            | 0.024        | 0.489                 | 0.214        | 0.096            | 0.100            | 0.719                 | 0.305        | 0.218            | 0.004        | 0.291                 | 7.367        | 2.554        |
| N5             | 0.014        | 0.535        | 20.044        | 6.864        | 0.119            | 0.032        | 0.472                 | 0.208        | 0.097            | 0.102            | 0.704                 | 0.301        | 0.194            | 0.006        | 0.320                 | 4.542        | 1.623        |
| <b>Average</b> | <b>0.033</b> | <b>0.423</b> | <b>16.015</b> | <b>5.490</b> | <b>0.119</b>     | <b>0.038</b> | <b>0.439</b>          | <b>0.199</b> | <b>0.096</b>     | <b>0.107</b>     | <b>0.718</b>          | <b>0.307</b> | <b>0.243</b>     | <b>0.004</b> | <b>0.264</b>          | <b>3.284</b> | <b>1.184</b> |

TABLE A2. AVERAGE OF ABSOLUTE ERROR RESULTS OF APPLYING FICA ALGORITHM FOR ALL NODES WITH DIFFERENT DATASETS

| FICA           |             |          |          |        |                  |             |                       |       |                     |                     |                       |       |                  |            |                       |             |        |
|----------------|-------------|----------|----------|--------|------------------|-------------|-----------------------|-------|---------------------|---------------------|-----------------------|-------|------------------|------------|-----------------------|-------------|--------|
| Nodes          | Data1-AirQ  |          |          |        | Data2-ARHO       |             |                       |       | Data3-GSB           |                     |                       |       | Data4_Intel      |            |                       |             |        |
|                | Temperature | Humidity | Pressure | Mean   | Temperature (°C) | soilT_30cmc | Relative Humidity (%) | Mean  | A. Temperature (°C) | S. Temperature (°C) | Relative Humidity (%) | Mean  | Temperature (°C) | Voltage(v) | Relative Humidity (%) | Light (Lux) | Mean   |
| N1             | 3.193       | 3.631    | 152.485  | 53.103 | 2.206            | 1.198       | 2.152                 | 1.852 | 0.883               | 0.490               | 6.779                 | 2.717 | 1.771            | 0.017      | 2.209                 | 128.169     | 43.465 |
| N2             | 3.193       | 3.631    | 152.485  | 53.103 | 2.252            | 1.897       | 3.297                 | 2.482 | 1.458               | 1.644               | 10.950                | 4.684 | 1.004            | 0.029      | 2.623                 | 109.780     | 37.477 |
| N3             | 4.410       | 8.748    | 143.465  | 52.208 | 2.394            | 1.444       | 1.316                 | 1.718 | 1.088               | 0.509               | 9.897                 | 3.831 | 17.764           | 0.130      | 4.089                 | 115.559     | 39.926 |
| N4             | 3.004       | 3.234    | 170.109  | 58.782 | 2.063            | 0.827       | 1.250                 | 1.380 | 1.102               | 0.428               | 9.115                 | 3.548 | 1.290            | 0.040      | 1.901                 | 36.201      | 12.714 |
| N5             | 1.565       | 6.314    | 275.148  | 94.343 | 1.987            | 1.178       | 1.583                 | 1.583 | 1.568               | 1.178               | 10.622                | 4.456 | 1.173            | 0.015      | 1.915                 | 27.623      | 9.851  |
| <b>Average</b> | 3.073       | 5.112    | 178.738  | 62.308 | 2.180            | 1.309       | 1.920                 | 1.803 | 1.220               | 0.850               | 9.473                 | 3.847 | 4.601            | 0.046      | 2.547                 | 83.467      | 28.687 |

TABLE A3. AVERAGE OF ABSOLUTE ERROR RESULTS OF APPLYING NNF ALGORITHM FOR ALL NODES WITH DIFFERENT DATASETS

| NNF            |             |          |          |        |                  |             |                       |       |                     |                     |                       |       |                  |            |                       |             |       |
|----------------|-------------|----------|----------|--------|------------------|-------------|-----------------------|-------|---------------------|---------------------|-----------------------|-------|------------------|------------|-----------------------|-------------|-------|
| Nodes          | Data1-AirQ  |          |          |        | Data2-ARHO       |             |                       |       | Data3-GSB           |                     |                       |       | Data4_Intel      |            |                       |             |       |
|                | Temperature | Humidity | Pressure | Mean   | Temperature (°C) | soilT_30cmc | Relative Humidity (%) | Mean  | A. Temperature (°C) | S. Temperature (°C) | Relative Humidity (%) | Mean  | Temperature (°C) | Voltage(v) | Relative Humidity (%) | Light (Lux) | Mean  |
| N1             | 0.000       | 16.472   | 4.318    | 6.930  | 0.000            | 14.270      | 1.768                 | 5.346 | 0.000               | 1.518               | 2.860                 | 1.459 | 0.000            | 2.190      | 0.011                 | 0.000       | 0.734 |
| N2             | 0.000       | 20.839   | 5.614    | 8.818  | 0.000            | 13.882      | 2.839                 | 5.574 | 0.000               | 1.882               | 3.242                 | 1.708 | 0.000            | 7.855      | 0.035                 | 0.000       | 2.630 |
| N3             | 0.000       | 25.032   | 5.963    | 10.332 | 0.000            | 14.229      | 2.449                 | 5.559 | 0.000               | 1.470               | 2.359                 | 1.276 | 0.000            | 2.256      | 0.045                 | 0.000       | 0.767 |
| N4             | 0.000       | 23.722   | 5.138    | 9.620  | 0.000            | 13.814      | 1.261                 | 5.025 | 0.000               | 1.340               | 1.800                 | 1.047 | 0.000            | 1.323      | 0.034                 | 0.000       | 0.452 |
| N5             | 0.000       | 18.361   | 3.503    | 7.288  | 0.000            | 14.688      | 1.567                 | 5.418 | 0.000               | 1.456               | 2.184                 | 1.213 | 0.000            | 1.381      | 0.009                 | 0.000       | 0.463 |
| <b>Average</b> | 0.000       | 20.885   | 4.907    | 8.597  | 0.000            | 14.177      | 1.977                 | 5.384 | 0.000               | 1.533               | 2.489                 | 1.341 | 0.000            | 3.001      | 0.027                 | 0.000       | 1.009 |

TABLE A4. AVERAGE OF ABSOLUTE ERROR RESULTS OF APPLYING NNTS ALGORITHM FOR ALL NODES WITH DIFFERENT DATASETS

| NNTS           |             |          |          |        |                  |             |                       |       |                     |                     |                       |       |                  |            |                       |             |       |
|----------------|-------------|----------|----------|--------|------------------|-------------|-----------------------|-------|---------------------|---------------------|-----------------------|-------|------------------|------------|-----------------------|-------------|-------|
| Nodes          | Data1-AirQ  |          |          |        | Data2-ARHO       |             |                       |       | Data3-GSB           |                     |                       |       | Data4_Intel      |            |                       |             |       |
|                | Temperature | Humidity | Pressure | Mean   | Temperature (°C) | soilT_30cmc | Relative Humidity (%) | Mean  | A. Temperature (°C) | S. Temperature (°C) | Relative Humidity (%) | Mean  | Temperature (°C) | Voltage(v) | Relative Humidity (%) | Light (Lux) | Mean  |
| N1             | 0.000       | 14.871   | 4.838    | 6.570  | 0.000            | 15.571      | 1.860                 | 5.810 | 0.000               | 2.840               | 1.682                 | 1.507 | 0.000            | 2.295      | 0.013                 | 0.000       | 0.769 |
| N2             | 0.000       | 36.685   | 17.200   | 17.962 | 0.000            | 14.526      | 2.835                 | 5.787 | 0.000               | 2.464               | 1.476                 | 1.313 | 0.000            | 9.455      | 0.280                 | 0.000       | 3.245 |
| N3             | 0.000       | 25.774   | 10.295   | 12.023 | 0.000            | 14.771      | 2.516                 | 5.762 | 0.000               | 2.222               | 1.533                 | 1.252 | 0.000            | 4.609      | 0.370                 | 0.000       | 1.660 |
| N4             | 0.000       | 19.630   | 5.119    | 8.250  | 0.000            | 14.295      | 1.200                 | 5.165 | 0.000               | 1.837               | 1.516                 | 1.118 | 0.000            | 3.719      | 0.038                 | 0.000       | 1.252 |
| N5             | 0.000       | 18.334   | 3.423    | 7.252  | 0.000            | 15.152      | 1.640                 | 5.597 | 0.000               | 2.272               | 1.617                 | 1.296 | 0.000            | 1.385      | 0.009                 | 0.000       | 0.465 |
| <b>Average</b> | 0.000       | 23.059   | 8.175    | 10.411 | 0.000            | 14.863      | 2.010                 | 5.624 | 0.000               | 2.327               | 1.565                 | 1.297 | 0.000            | 4.293      | 0.142                 | 0.000       | 1.478 |

TABLE A5. AVERAGE OF ABSOLUTE ERROR RESULTS OF APPLYING LRMV ALGORITHM FOR ALL NODES WITH DIFFERENT DATASETS

| LRMV           |             |          |          |        |                  |             |                       |       |                     |                     |                       |       |                  |            |                       |             |       |
|----------------|-------------|----------|----------|--------|------------------|-------------|-----------------------|-------|---------------------|---------------------|-----------------------|-------|------------------|------------|-----------------------|-------------|-------|
| Nodes          | Data1-AirQ  |          |          |        | Data2-ARHO       |             |                       |       | Data3-GSB           |                     |                       |       | Data4_Intel      |            |                       |             |       |
|                | Temperature | Humidity | Pressure | Mean   | Temperature (°C) | soilT_30cmc | Relative Humidity (%) | Mean  | A. Temperature (°C) | S. Temperature (°C) | Relative Humidity (%) | Mean  | Temperature (°C) | Voltage(v) | Relative Humidity (%) | Light (Lux) | Mean  |
| N1             | 0.000       | 69.680   | 0.000    | 23.227 | 0.000            | 1.212       | 0.000                 | 0.404 | 0.000               | 5.211               | 0.000                 | 1.737 | 0.000            | 1.771      | 0.017                 | 0.000       | 0.596 |
| N2             | 0.000       | 69.680   | 0.000    | 23.227 | 0.000            | 2.132       | 0.000                 | 0.711 | 0.000               | 1.933               | 0.000                 | 0.644 | 0.000            | 1.004      | 0.029                 | 0.000       | 0.345 |
| N3             | 0.000       | 133.045  | 0.000    | 44.348 | 0.000            | 2.615       | 0.000                 | 0.872 | 0.000               | 1.460               | 0.000                 | 0.487 | 0.000            | 17.764     | 0.130                 | 0.000       | 5.965 |
| N4             | 0.000       | 25.408   | 0.000    | 8.469  | 0.000            | 0.845       | 0.000                 | 0.282 | 0.000               | 1.411               | 0.000                 | 0.470 | 0.000            | 1.290      | 0.040                 | 0.000       | 0.444 |
| N5             | 0.000       | 4.274    | 0.000    | 1.425  | 0.000            | 1.278       | 0.000                 | 0.426 | 0.000               | 2.587               | 0.000                 | 0.862 | 0.000            | 1.173      | 0.015                 | 0.000       | 0.396 |
| <b>Average</b> | 0.000       | 60.417   | 0.000    | 20.139 | 0.000            | 1.616       | 0.000                 | 0.539 | 0.000               | 2.520               | 0.000                 | 0.840 | 0.000            | 4.601      | 0.046                 | 0.000       | 1.549 |

# A Global Survey of Technological Resources and Datasets on COVID-19

Manoj Muniswamaiah, Tilak Agerwala, Charles C. Tappert  
Seidenberg School of CSIS, Pace University, White Plains, New York

**Abstract**—The application and successful utilization of technological resources in developing solutions to health, safety, and economic issues caused by COVID-19 indicate the importance of technology in curbing COVID-19. Also, the medical field has had to race against time to develop and distribute the COVID-19 vaccine. This endeavour became successful with the vaccines created and approved in less than a year, a feat in medical history. Currently, much work is being done on data collection, where all significant factors impacting the disease are recorded. These factors include confirmed cases, death rates, vaccine rates, hospitalization data, and geographic regions affected by the pandemic. Continued research and use of technological resources are highly recommendable—the paper surveys list of packages, applications and datasets used to analyse COVID-19.

**Keywords**—Vaccination; hospitalization; confirmed cases; datasets; data science; COVID-19

## I. INTRODUCTION

COVID-19 pandemic has affected the world; data is being collected by agencies, organizations, institutions, and other bodies that are keen on providing insights [391]. Data collection includes conducting case surveillance to gather data on demographics, clinical factors, epidemiologic characteristics, illness course, care, and history on exposure and contact. This data is needed to assess where, when, and

who are most affected by the pandemic. The data available on COVID-19 is used by researchers in the medical field in evaluating different aspects of the virus. Statistical analysis tools estimate factors concerning the virus infectiousness obtained in growth rate and doubling time. Epidemiological models are used to group individuals based on their demographic data and apply mathematical formulas to find virus characteristics. Using machine learning, the time series prediction model is proposed to obtain the curve and forecast the epidemic's tendencies.

As a part of tackling the pandemic, different institutions, governments, organizations, and individuals have developed and adopted technological resources to manage the pandemic and support adherence to containment measures. Most of the technologies developed have utilized R programming, Python, Java, Kotlin, JavaScript, among other resources. Mobile applications were developed to help in contact tracing, notifications, and alerts to users if they interacted with a person infected. Dashboards have been used in visualizing COVID-19 cases across the world. This paper presents a survey of the technological resources and datasets been used by industry and academia to combat COVID-19.

Table I provide the name of the application; the details of the developer, institution, or academia; the summary of the application, its web link, and codebase link.

TABLE I. TECHNOLOGICAL RESOURCES SUMMARY TABLE

| Application                                 | Developer/industry/ Academia/University details | Application summary                                                                                                                                                                                                                                                                 | Application weblink                                                                                                                 | Application code base link                                                                                                    | References  |
|---------------------------------------------|-------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|-------------|
| Coronavirus tracker                         | Developed by John Coene.                        | Coronavirus tracker is an R Shiny app that tracks the spread of the coronavirus, based on three data sources, including John Hopkins, Weixin, and DXY Data. The app summarizes the coronavirus statistics such as deaths, confirmed, recovered, and suspected cases on a dashboard. | <a href="https://www.coronatracker.com/">https://www.coronatracker.com/</a> [1]                                                     | <a href="https://github.com/JohnCoene/coronavirus.git">https://github.com/JohnCoene/coronavirus.git</a> [2]                   | [1] [2] [3] |
| COVID-19 Global Cases                       | Developed by Christoph Schoenenberger.          | COVID-19 Global cases are a shiny app that displays the recent Covid-19 developments via key figures, plots, a map, and summary tables.                                                                                                                                             | <a href="https://chschoenenberger.shinyapps.io/covid19_dashboard/">https://chschoenenberger.shinyapps.io/covid19_dashboard/</a> [4] | <a href="https://github.com/chschoenenberger/covid19_dashboard">https://github.com/chschoenenberger/covid19_dashboard</a> [5] | [3] [4] [5] |
| The 2019-20 Coronavirus Pandemic A timeline | Developed by Nico Hahn                          | Visualization of Covid-19 cases is a shiny application that uses leaflet, plotly, and data from Johns Hopkins University to visualize the novel coronavirus outbreak and show data for the entire world or particular countries.                                                    | <a href="https://nicohahn.shinyapps.io/covid19/">https://nicohahn.shinyapps.io/covid19/</a> [6]                                     | <a href="https://github.com/nicoFhahn/covid_shiny">https://github.com/nicoFhahn/covid_shiny</a> [7]                           | [3] [6] [7] |

|                                                          |                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                                          |                                                                                                                                            |                                                                                                                                                                                |                |
|----------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| Modeling COVID-19 Spread vs Healthcare Capacity          | Developed by Dr. Alison Hill from Johns Hopkins University                                                                                                                                                                                  | This application utilizes the epidemiological model based on the classic SEIR model to define the Covid-19 spread and clinical progression. The application provides different infection trajectories, clinical interventions to curb transmission, and a comparison to the current healthcare capacity. | <a href="https://alhill.shinyapps.io/COVID19seir/">https://alhill.shinyapps.io/COVID19seir/</a> [8]                                        | <a href="https://github.com/alsnhll/SEIR_COVID19">https://github.com/alsnhll/SEIR_COVID19</a> [9]                                                                              | [3] [8] [9]    |
| COVID-19 Data Visualization Platform                     | Developed by Shubhram Pandey.                                                                                                                                                                                                               | This is a shiny app that provides an elaborate visualization of the impact of Covid-19 across the globe. The application also applies natural language processing from Twitter to provide sentiment analysis.                                                                                            | <a href="https://shubhrampandey.shinyapps.io/coronaVirusViz/">https://shubhrampandey.shinyapps.io/coronaVirusViz/</a> [10]                 | <a href="https://github.com/shubhrampandey/coronaVirus-dataViz">https://github.com/shubhrampandey/coronaVirus-dataViz</a> [11]                                                 | [3] [10] [11]  |
| Coronavirus 10-day forecast                              | This application was developed by Spatial Ecology and Evolution Lab (SpEEL) from the University of Melbourne.                                                                                                                               | It is a shiny app that provides a ten-day forecast of likely coronavirus cases by country, giving individuals a sense of how the Covid-19 is spreading or progressing.                                                                                                                                   | <a href="https://covid19forecast.science.unimelb.edu.au/">https://covid19forecast.science.unimelb.edu.au/</a> [12]                         | <a href="https://github.com/benflips/nCovForecast">https://github.com/benflips/nCovForecast</a> [13]                                                                           | [3] [12] [13]  |
| Coronavirus (COVID-19) across the world                  | This application was developed by Anisa Dhana.                                                                                                                                                                                              | It is a shiny app that uses a map visualization of cases confirmed to monitor the spread of Covid-19 across the world and graphs to visualize the growth of the disease.                                                                                                                                 | <a href="https://dash.data-scienceplus.com/covid19/">https://dash.data-scienceplus.com/covid19/</a> [14]                                   | <a href="https://github.com/CSSEGISandData/COVID-19">https://github.com/CSSEGISandData/COVID-19</a> [15]                                                                       | [16] [14] [15] |
| COVID-19 outbreak                                        | Dr. Thibaut Fabacher developed this application in collaboration with the department of Public Health of the Strasbourg University Hospital and the Laboratory of Biostatistics and Medical Informatics of the Strasbourg Medicine Faculty. | The application displays an interactive map that indicates the worldwide monitoring of Covid-19 infection. The main area of focus of the app is on the evolution of the number of Covid-19 cases per country for a given period.                                                                         | <a href="https://thibautfabacher.shinyapps.io/covid-19/">https://thibautfabacher.shinyapps.io/covid-19/</a> [17]                           | <a href="https://github.com/DrFabacher/Corona">https://github.com/DrFabacher/Corona</a> [18]                                                                                   | [3] [17] [18]  |
| Corona trajectories                                      | This application was developed by André Calero Valdez, from RWTH Aachen University.                                                                                                                                                         | The application uses two graphs to compare the number of confirmed cases and the deaths from Covid-19 with the country's trajectories. The application also allows users to compare the case number and growth rate of the Covid-19 pandemic per country using a table.                                  | <a href="https://andrevalerovaldez.shinyapps.io/CovidTimeSeriesTest/">https://andrevalerovaldez.shinyapps.io/CovidTimeSeriesTest/</a> [19] | <a href="https://github.com/Sumidu/covid19shiny">https://github.com/Sumidu/covid19shiny</a> [20]                                                                               | [3] [19] [20]  |
| Flatten the curve                                        | Tinu Schneider developed                                                                                                                                                                                                                    | this application.<br>In an interactive way, the app illustrates the different scenarios behind the #FlattenTheCurve message.                                                                                                                                                                             | <a href="https://tinu.shinyapps.io/Flatten_the_Curve/">https://tinu.shinyapps.io/Flatten_the_Curve/</a> [21]                               | <a href="https://github.com/tinuschneider/Flatten_the_Curve">https://github.com/tinuschneider/Flatten_the_Curve</a> [22]                                                       | [3] [21] [22]  |
| Explore the Spread of Covid-19                           | Joachim Gassen developed this application,                                                                                                                                                                                                  | The application allows users to visualize confirmed, recovered cases and reported deaths for several countries via one summary graph.                                                                                                                                                                    | <a href="https://jgassen.shinyapps.io/tidyCovid19/">https://jgassen.shinyapps.io/tidyCovid19/</a> [23]                                     | <a href="https://statsandr.com/blog/top-resources-on-covid-19-coronavirus/#tidycovid19">https://statsandr.com/blog/top-resources-on-covid-19-coronavirus/#tidycovid19</a> [24] | [3] [23] [24]  |
| COVID-19                                                 | Sebastian Engel-Wolf developed the application                                                                                                                                                                                              | The application visualizes elegantly collected Covid-19 data, including the confirmed cases, Maximum time of exponential growth in a row, deaths, populations, and Confirmed cases on 100,000 inhabitants, exponential growth, and the population.                                                       | <a href="https://sebastianwolf.shinyapps.io/Corona-Shiny/">https://sebastianwolf.shinyapps.io/Corona-Shiny/</a> [25]                       | <a href="https://github.com/zappingseb/coronashiny">https://github.com/zappingseb/coronashiny</a> [26]                                                                         | [3] [25] [26]  |
| Simulation tool. COVID-19 epidemic in Togo - West Africa | Dr. Kankoé Sallah developed this application.                                                                                                                                                                                               | Uses SEIR metapopulation model with mobility between catchment areas to describe the country-level spread of COVID-19 and the impact of interventions in Togo, West Africa.                                                                                                                              | <a href="https://c2m-africa.shinyapps.io/togo-covid-shiny/">https://c2m-africa.shinyapps.io/togo-covid-shiny/</a> [27]                     |                                                                                                                                                                                | [3] [27]       |
| Animating                                                | Nathan Chaney                                                                                                                                                                                                                               | This application indicates a map animation of                                                                                                                                                                                                                                                            |                                                                                                                                            | <a href="https://www.nathanchaney.com/">https://www.nathanchaney.com/</a>                                                                                                      | [30] [29]      |

|                                                                      |                                                                                                                                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                                    |                                                                                                                                                                                                                  |                |
|----------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------|
| COVID-19 hotspots over time                                          | developed this application                                                                                                                | new Covid-19 cases in the U.S.A measured in a 7-day rolling average.                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                    | <a href="#">com/</a> [29]                                                                                                                                                                                        |                |
| Covid-19-prediction                                                  | This application was developed by Manuel Oviedo and Manuel Febrero of Modesty research group of the University of Santiago de Compostela. | The application is a shiny app that provides a 5-day horizon prediction growth rate of Covid-19 using the evolution during the past 15-day growth rate. The prediction is performed using three functional regression models fitted and estimated on available data. Apart from the prediction values, the app provides an interactive table and plot for the expected number of accumulated cases and new daily confirmed and death cases.                  | <a href="http://modesty.us.es:3838/covid19prediction/">http://modesty.us.es:3838/covid19prediction/</a> [33]                       | <a href="https://github.com/armimpdm/Covid-19-prediction">https://github.com/armimpdm/Covid-19-prediction</a> [34]                                                                                               | [3] [33] [34]  |
| Healthcare worker deaths from novel Coronavirus (COVID-19) in the US | Jonathan Gross developed this application                                                                                                 | The application is a shiny app that visualizes the U.S. health workers' deaths from Covid-19 reported on media outlets or news. The application is developed using R code with a map on the main page using Leaflet with tabs for additional graphs, including time series, histograms, and bar charts.                                                                                                                                                      | <a href="https://jontheepi.shinyapps.io/hcwcoronavirus/">https://jontheepi.shinyapps.io/hcwcoronavirus/</a> [35]                   | <a href="https://github.com/jontheepi/hcwcoronavirus">https://github.com/jontheepi/hcwcoronavirus</a> [36]                                                                                                       | [3] [35] [36]  |
| Covid-19 Hospitalizations in Belgium                                 | Jean-Michel Bodart developed this application                                                                                             | The dashboard indicates the hospitalizations related to Covid-19 in Belgium by province and region.                                                                                                                                                                                                                                                                                                                                                          | <a href="https://rpubs.com/JMBodart/Covid19-hosp-be">https://rpubs.com/JMBodart/Covid19-hosp-be</a> [37]                           | <a href="https://github.com/jmbo1190/Covid19">https://github.com/jmbo1190/Covid19</a> [38]                                                                                                                       | [3] [37] [38]  |
| Covidminer                                                           | This shiny app was developed by the Rensselaer Institute for Data Exploration and Applications                                            | The application indicates the regional differences in determinants, medications, and outcome of the Covid-19 pandemic across the United but with a specific focus on New York.                                                                                                                                                                                                                                                                               | <a href="https://covidminer.idea.rpi.edu/">https://covidminer.idea.rpi.edu/</a> [39]                                               | <a href="https://github.com/TheRensselaerIDEA/COVIDMINER">https://github.com/TheRensselaerIDEA/COVIDMINER</a> [40]                                                                                               | [41] [39] [40] |
| COVID-19 Canada Data Explorer                                        | Petr Baranovskiy developed this application.                                                                                              | The application is a shiny app that analyses the official covid-19 dataset from the government of Canada and outputs the several indicators associated with the Covid-19 pandemic in the country.                                                                                                                                                                                                                                                            | <a href="https://dataenthusiast.ca/apps/covid-ca/">https://dataenthusiast.ca/apps/covid-ca/</a> [42]                               | <a href="https://milano-r.github.io/erum2020-covid-contest/petr-baranovskiy-covid-ca-data-explorer.html">https://milano-r.github.io/erum2020-covid-contest/petr-baranovskiy-covid-ca-data-explorer.html</a> [43] | [3] [42][43]   |
| PAGTANA: Philippine COVID-19 Case Forecasting Web Application        | This application was developed by Jamal Kay Rogers and Yvonne Grace Arandela.                                                             | It is a shiny app that provides a 5-day forecast of Covid-19 cases in the Philippines include the confirmed new cases of infections, confirmed deaths, and recovery rate. Apart from forecasting, the application utilizes plots to visualize the disease's ten-day forecasts and the accumulated and confirmed data. The data used in this application is obtained from the Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE). | <a href="https://jamalrogersapp.shinyapps.io/tforecast/">https://jamalrogersapp.shinyapps.io/tforecast/</a> [44]                   | <a href="https://github.com/fsmosca/COVID-19-PH-dataset">https://github.com/fsmosca/COVID-19-PH-dataset</a> [45]                                                                                                 | [3] [44] [45]  |
| COVID-19 Case & Death Report Number Corrector                        | Matt Maciejewski developed this application.                                                                                              | This shiny application is developed and aligned to make corrections of underreported Covid-19 cases and death. The application applies a multiplicative estimator for total deaths and cases regarding the base country to perform this role.                                                                                                                                                                                                                | <a href="https://pharmhax.shinyapps.io/covid-corrector-shiny/">https://pharmhax.shinyapps.io/covid-corrector-shiny/</a> [46]       | <a href="https://github.com/pharmhax/covid19-corrector">https://github.com/pharmhax/covid19-corrector</a> [47]                                                                                                   | [48] [46] [47] |
| COVID19 forecast                                                     | Carlos Catania developed this application                                                                                                 | This application applies the SEIR model to forecast the spread of Covid 19 in various European and South American countries.                                                                                                                                                                                                                                                                                                                                 | <a href="https://harpomaxx.shinyapps.io/covid19/">https://harpomaxx.shinyapps.io/covid19/</a> [49]                                 | <a href="https://github.com/harpomaxx/COVID19">https://github.com/harpomaxx/COVID19</a> [50]                                                                                                                     | [3] [49] [50]  |
| Trafford Covid-19 monitor                                            | This is a shiny application developed by Trafford Data Lab                                                                                | The application provides trends in confirmed coronavirus cases in Trafford.                                                                                                                                                                                                                                                                                                                                                                                  | <a href="https://trafforddatalab.shinyapps.io/trafford-covid-19/">https://trafforddatalab.shinyapps.io/trafford-covid-19/</a> [54] | <a href="https://github.com/traffordDataLab/trafford-covid-19">https://github.com/traffordDataLab/trafford-covid-19</a> [55]                                                                                     | [53] [54] [55] |
| Covid-19 Bulletin Board                                              | Wei Su developed this application                                                                                                         | The dashboard indicates the real-time Covid-19 visualization of the various covid-19 indicators in Japan, including the confirmed cases, hospital discharge and deaths, positive confirmed, and PCR test.                                                                                                                                                                                                                                                    | <a href="https://covid-2019.live/en/">https://covid-2019.live/en/</a> [56]                                                         | <a href="https://github.com/swsoyee/2019-ncov-japan">https://github.com/swsoyee/2019-ncov-japan</a> [57]                                                                                                         | [3] [56] [57]  |
| Covid-19 Statistics                                                  | Carl Sansfaçon developed this                                                                                                             | It is a WordPress plugin that applies the R {ggplot2} graphics with ARIMA forecast and                                                                                                                                                                                                                                                                                                                                                                       | <a href="http://moduloinfo.ca/wordpress/">http://moduloinfo.ca/wordpress/</a> [58]                                                 | <a href="https://plugins.trac.wordpress.org/browser/covid-19-">https://plugins.trac.wordpress.org/browser/covid-19-</a>                                                                                          | [3] [58] [59]  |

|                                                                         |                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                            |                                                                                                                                                                          |                          |
|-------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------|
| Displayer                                                               | application                                                                                                        | PHP coding to display or visualize the confirmed new cases of Covid-19 infection, deaths, and recovered cases in various countries. The data used in this application is sourced from the COVID-19 Data Repository by the Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University.                                                                                                                                                                                                                                                                  |                                                                                                                            | <a href="#">statistics-displayer/</a> [59]                                                                                                                               |                          |
| CoronaMapper                                                            | This application was developed by Peter Gruber and Paolo Montemurro supported by OxyLabs.                          | The application visualizes the four-day average growth indicator of Covid-19 to indicate how the disease evolves after filtering out the noise. The visualizations are both interactive and intuitive.                                                                                                                                                                                                                                                                                                                                                                    | <a href="http://coronamapper.com/">http://coronamapper.com/</a> [60]                                                       | <a href="https://github.com/JayWelsch/coronamapper">https://github.com/JayWelsch/coronamapper</a> [61]                                                                   | [3] [60] [61]            |
| CoronaDash                                                              | This is a shiny app developed by Peter Laurinec.                                                                   | This application applies visualization and data mining techniques in R to compare Covid-19 statistics for different countries. The Covid-19 statistics displayed are obtained by using exponential smoothing model to extrapolate total confirmed cases; creating death trajectories; using dendrogram and table of clusters averages to create a multidimensional clustering; developing aggregated views of the entire world; and applying hierarchical clustering to compare the Covid-19 case between countries.                                                      | <a href="https://petolau.shinyapps.io/coronadash/">https://petolau.shinyapps.io/coronadash/</a> [62]                       | <a href="https://github.com/PetoLau/CoronaDash">https://github.com/PetoLau/CoronaDash</a> [63]                                                                           | [3] [62] [63]            |
| Covidfrance                                                             | This is a shiny app developed by Guillaume Pressiat                                                                | The application indicates the changes in the number of Covid-19 deaths and recoveries, hospitalization, and intensive care units by the department in France                                                                                                                                                                                                                                                                                                                                                                                                              | <a href="https://guillaumepressiat.shinyapps.io/covidfrance/">https://guillaumepressiat.shinyapps.io/covidfrance/</a> [64] | <a href="https://gist.github.com/GuilLaumePressiat/0e3658624e42f763e3e6a67df92bc6c5">https://gist.github.com/GuilLaumePressiat/0e3658624e42f763e3e6a67df92bc6c5</a> [65] | [3] [64] [65]            |
| COVID-19 Tracker                                                        | Dr Magda Bucholc developed this application from Ulster University                                                 | The application reports the number of reported Covid-19 cases at the local government district in Northern Ireland and county level across Ireland based on gender and growth rate.                                                                                                                                                                                                                                                                                                                                                                                       | <a href="https://nicovidtracker.org/">https://nicovidtracker.org/</a> [66]                                                 | <a href="https://github.com/YouGov-Data/covid-19-tracker">https://github.com/YouGov-Data/covid-19-tracker</a> [67]                                                       | [68] [66] [67]           |
| WHO COVID-19 Explorer                                                   | This application was developed by the World Health Organization (WHO)                                              | This application provides timely updated data visualizations of Covid-19 cases, including confirmed cases and deaths by region and country.                                                                                                                                                                                                                                                                                                                                                                                                                               | <a href="https://worldhealthorg.shinyapps.io/covid/">https://worldhealthorg.shinyapps.io/covid/</a> [71]                   | <a href="https://github.com/WorldHealthOrganization/app">https://github.com/WorldHealthOrganization/app</a> [72]                                                         | [3] [71] [72]            |
| COVID-19 Scenario Analysis Tool                                         | The MRC Centre developed this application for Global Infectious Disease Analysis from the Imperial College London. | This application applies the squire R package to illustrate the Covid-19 pandemic trajectories, R_t & R_eff measures, and healthcare demand for different countries over time.                                                                                                                                                                                                                                                                                                                                                                                            | <a href="https://www.covid-sim.org/v6.20210915/">https://www.covid-sim.org/v6.20210915/</a> [73]                           | <a href="https://github.com/mrc-ide/squire">https://github.com/mrc-ide/squire</a> [74]                                                                                   | [3] [73] [74]            |
| Coronavirus Package                                                     | Rami Krispin developed this R package                                                                              | This package provides a clean dataset of the Covid-19 pandemic and analytics, including the daily summary of the pandemic cases by state. The dataset is collected from the John Hopkins database.                                                                                                                                                                                                                                                                                                                                                                        |                                                                                                                            | <a href="https://ramikrispin.github.io/coronavirus/">https://ramikrispin.github.io/coronavirus/</a> [75]                                                                 | [3] [75]                 |
| District Health Information Software (DHIS2)                            | The University of Oslo developed this application                                                                  | The District Health Information Software has specific digital packages for Covid-19 that support the pandemic's surveillance and response activities.                                                                                                                                                                                                                                                                                                                                                                                                                     | <a href="https://www.dhis2.org/">https://www.dhis2.org/</a> [76]                                                           | <a href="https://github.com/dhis2/dhis2-covid19-doc">https://github.com/dhis2/dhis2-covid19-doc</a> [79]                                                                 | [80] [76] [77] [78] [79] |
| Surveillance, Outbreak Response Management and Analysis System (SORMAS) | Helmholtz Centre for Infection Research developed this system.                                                     | The system performs the Covid-19 specific functions that are classified into aggregates and case-based functions. The aggregate functions include line listing, import, and export of data in CSV format; standard reporting of covid-19 cases including confirmed cases, deaths, and recoveries; and statistical analysis based on the reports provided by charts, maps, and graphs. The case-based functions include contact tracing, laboratory sample management, port of entry reporting, vaccination campaign, follow-up visit, and enrolling and tracing patients. | <a href="https://sormas.org/">https://sormas.org/</a> [81]                                                                 | <a href="https://github.com/hzi-braunschweig/SORMAS-Project">https://github.com/hzi-braunschweig/SORMAS-Project</a> [82]                                                 | [80] [81] [82]           |

|                                                          |                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                          |                                                                                                                                                                    |                     |
|----------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Go.Data                                                  | This application was developed by WHO in collaboration with partners in the Global Outbreak Alert and Response Network (GOARN).                                                                                                                                                                                                                                                  | Since the outbreak of Covid-19 began, metadata packages have been prepared that match the most recent WHO Surveillance Guidance, including uniformity with all core metadata gathered as part of WHO Case Reporting Forms transmitted to COVID-MART / X-MART on a daily and weekly basis. If requested, this allows for streamlined IDSR reporting for countries. Other expanded metadata packages, such as the COVID First Few Hundred Cases (FFX) Protocol and the Unity Studies for HealthCare Workers, are available for countries conducting more extensive data collection or research inquiries. | <a href="https://www.who.int/godata">https://www.who.int/godata</a> [83]                                                 | <a href="https://github.com/godata-who/godata">https://github.com/godata-who/godata</a> [84]                                                                       | [80] [83] [84]      |
| Epi Info                                                 | This application was developed by Centres for Disease Control and Prevention (CDC)                                                                                                                                                                                                                                                                                               | The Covid-19 specific functions include the development of COVID-19 Case Surveillance Forms that are customized for country, region, and local requirements. Epi Info is also applied in Covid-19 outbreak investigations, the development of small to mid-sized disease surveillance systems, the analysis, visualization, and reporting (AVR) components of larger systems, and continuing education in epidemiology and public health analytic methods at public health schools around the world.                                                                                                    | <a href="https://www.cdc.gov/epiinfo/support/downloads.html">https://www.cdc.gov/epiinfo/support/downloads.html</a> [85] | <a href="https://github.com/Epi-Info/Epi-Info-Community-Edition">https://github.com/Epi-Info/Epi-Info-Community-Edition</a> [86]                                   | [80] [85] [86]      |
| Open Data Kit (ODK)                                      | This application was developed by Get ODK, an organization majoring in data collection.                                                                                                                                                                                                                                                                                          | ODK software is being employed in the COVID-19 response for disease surveillance, fast diagnostics, and vaccine trials.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 | <a href="https://getodk.org/software/">https://getodk.org/software/</a> [87]                                             | <a href="https://github.com/getodk/collect">https://github.com/getodk/collect</a> [88]                                                                             | [80] [87] [88]      |
| CommCare                                                 | This software was developed by Dimagi, a firm providing digital data solutions                                                                                                                                                                                                                                                                                                   | Dimagi created a set of pre-built COVID-19 template applications to help organizations and governments with their continuing COVID-19 response operations.                                                                                                                                                                                                                                                                                                                                                                                                                                              | <a href="https://www.dimagi.com/covid-19/">https://www.dimagi.com/covid-19/</a> [89]                                     | <a href="https://github.com/dimagi/commcare-hq">https://github.com/dimagi/commcare-hq</a> [90]                                                                     | [80] [89] [90]      |
| KoboToolbox                                              | This software was developed by the Harvard Humanitarian Initiative, an organization working on the research and education of communities.                                                                                                                                                                                                                                        | KoBoToolbox is a data collecting tool.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | <a href="https://www.kobotoolbox.org/">https://www.kobotoolbox.org/</a> [91][92]                                         | <a href="https://github.com/kobotoolbox">https://github.com/kobotoolbox</a> [93]                                                                                   | [80] [91] [92] [93] |
| Fast automated detection of COVID-19 from medical images | This application was developed by Shuang Liang & Huixiang Liu from School of Automation and Electrical Engineering, University of Science and Technology Beijing; Yu Gu from School of Automation, Guangdong University of Petrochemical Technology, Maoming; Xiuhua Guo, Zhiyuan Wu, Mengyang Liu & Lixin Tao<br><br>From the Department of Epidemiology and Health Statistics, | The software utilizes deep learning framework (neural network) that identifies COVID-19 from medical images.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |                                                                                                                          | <a href="https://github.com/SHERLOCKLS/Detection-of-COVID-19-from-medical-images">https://github.com/SHERLOCKLS/Detection-of-COVID-19-from-medical-images</a> [94] | [95] [94]           |

|                                                          |                                                                                                                                                                                                        |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                   |                                                                                                                         |                         |
|----------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|-------------------------|
|                                                          | School of Public Health, Capital Medical University, Beijing, China; and Hongjun Li & Li from Beijing Youan Hospital, Capital Medical University, Beijing, China.                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                   |                                                                                                                         |                         |
| The Oxford Covid-19 Government Response Tracker (OxCGRT) | This application was developed by the Blavatnik School of Government.                                                                                                                                  | The Oxford Covid-19 Government Response Tracker (OxCGRT) compiles systematic data on policy responses taken by countries to combat COVID-19. Since January 1, 2020, the various policy reactions have tracked over 180 nations and are categorized into 23 indicators, such as school closures, travel restrictions, and vaccination policies. These policies are scored on a scale to represent the magnitude of government intervention, and the results are compiled into a set of policy indices. The data can improve attempts to combat the epidemic by allowing decision-makers and citizens to understand government responses uniformly. | <a href="https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker">https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker</a> [96]                              | <a href="https://github.com/OxCGRT/covid-policy-tracker">https://github.com/OxCGRT/covid-policy-tracker</a> [97]        | [98] [96][97]           |
| COVID-19 Situazione Italia                               | This application was developed by the Department of Civil Protection (Dipartimento della Protezione Civile) Angelo Borrelli, Italy.                                                                    | This application provides updated Covid-19 data and visualizations for Italy, including new confirmed infections, total confirmed infections, new confirmed deaths, total confirmed deaths, and recovered cases. The data and visualizations are provided for the whole country and the regions.                                                                                                                                                                                                                                                                                                                                                  | <a href="http://arcg.is/C1unv">http://arcg.is/C1unv</a>                                                                                                                                                                           | <a href="https://github.com/pcm-dpc/COVID-19">https://github.com/pcm-dpc/COVID-19</a> [99]                              | [3] [99]                |
| Covid Mobile data                                        | This application was developed by COVID19 Mobility Task Force of the World Bank                                                                                                                        | The application uses the data from Mobile Network Operators (MNOs) to perform analytics.                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                                                                                   | <a href="https://github.com/worldbank/covid-mobile-data">https://github.com/worldbank/covid-mobile-data</a> [100]       | [3] [100]               |
| Radar Covid-19                                           | The Government of Spain developed this application                                                                                                                                                     | This application was designed to prevent the spread of Covid-19. The application anonymizes users if they have had any contact in the last 14 days with someone infected with Covid-19 via low-power Bluetooth technology.                                                                                                                                                                                                                                                                                                                                                                                                                        | <a href="https://radarcovid.gob.es/">https://radarcovid.gob.es/</a> [101]                                                                                                                                                         | <a href="https://github.com/RadarCOVID/radar-covid-android">https://github.com/RadarCOVID/radar-covid-android</a> [103] | [104] [101] [102],[103] |
| CovidSafe                                                | The University of Washington developed this application.                                                                                                                                               | The application was developed to help prevent the spread of Covid-19 by alerting users about highly relevant public health announcements, exposure to COVID-19 and to assist contact tracing without compromising users' privacy.                                                                                                                                                                                                                                                                                                                                                                                                                 | <a href="https://covidsafe.csis.washington.edu/">https://covidsafe.csis.washington.edu/</a> [105]                                                                                                                                 | <a href="https://github.com/CovidSafe">https://github.com/CovidSafe</a> [106]                                           | [3] [105] [106]         |
| Covid Alert                                              | Volunteers originally developed COVID Alert. The Canadian Digital Service is currently developing its repository.                                                                                      | This application was developed to slow down Covid-19 infections in Canada. The app notifies users if someone they were near in the past 14 days tells the app they tested positive.                                                                                                                                                                                                                                                                                                                                                                                                                                                               | <a href="https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html">https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html</a> [107] | <a href="https://github.com/cds-snc/covid-alert-app">https://github.com/cds-snc/covid-alert-app</a> [109]               | [110] [107] [108] [109] |
| erouska-android                                          | A team of volunteers initially developed this application. The application is currently developed and maintained by the Ministry of Health in collaboration with the National Agency for Communication | To combat the COVID-19 epidemic, the app alerts users at risk of spreading the virus. The software delivers guidance on how to minimize the spread of the epidemic based on the user's history of exposure to other potentially contagious users.                                                                                                                                                                                                                                                                                                                                                                                                 | <a href="https://erouska.cz/">https://erouska.cz/</a> [111]                                                                                                                                                                       | <a href="https://github.com/covid19cz/erouska-android">https://github.com/covid19cz/erouska-android</a> [112]           | [113] [111] [112]       |

|                          |                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                     |                                                                                                                           |                            |
|--------------------------|------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|----------------------------|
|                          | and Information Technologies (NAKIT) of the Czech Republic as part of the Smart Quarantine concept.                                |                                                                                                                                                                                                                                                                                                                                                                     |                                                                                                                                                                                                                                                                                                     |                                                                                                                           |                            |
| COVID-19 Dashboard       | Johns Hopkins University Centre developed this application for Systems Science and Engineering.                                    | This is the data repository for the Johns Hopkins University Centre for Systems Science and Engineering's 2019 Novel Coronavirus Visual Dashboard (JHU CSSE). The ESRI Living Atlas Team and the Johns Hopkins University Applied Physics Lab have also contributed to this project (JHU APL).                                                                      | <a href="https://www.arcgis.com/apps/opsdashboards/index.html#/bda7594740fd40299423467b48e9ecf6">https://www.arcgis.com/apps/opsdashboards/index.html#/bda7594740fd40299423467b48e9ecf6</a> [114]                                                                                                   | <a href="https://github.com/sidbannet/COVID-19_analysis">https://github.com/sidbannet/COVID-19_analysis</a> [116]         | [98] [114]<br>[115] [116]  |
| Corona-Warn-App          | This application was developed as an open-source app by SAP and Deutsche Telekom under the directive by the government of Germany. | The Corona-Warn-App was developed with the goal of preventing the spread of Covid-19. The app serves as a digital complement to distancing, hygiene, and wearing masks. Additionally, it provides a functionality to add a user's digital vaccination certificate to prove their vaccination status.                                                                | <a href="https://www.coronawarn.app/en/">https://www.coronawarn.app/en/</a> [117]                                                                                                                                                                                                                   | <a href="https://github.com/coronawarn-app/cwa-app-android">https://github.com/coronawarn-app/cwa-app-android</a> [118]   | [119] [117]<br>[118]       |
| TraceTogether            | The Singapore Government Technology Agency developed this application.                                                             | Through community-driven contact tracing, TraceTogether supports Singapore's efforts to combat the spread of COVID-19. One can use the app to see or display their COVID Health Status based on their immunization and test results.                                                                                                                                | <a href="https://www.tracetgether.gov.sg/">https://www.tracetgether.gov.sg/</a> [120]                                                                                                                                                                                                               | <a href="https://github.com/OpenTrace-Community">https://github.com/OpenTrace-Community</a> [121]                         | [122] [120]<br>[121]       |
| NZ COVID Tracer          | The New Zealand Ministry of Health developed this application                                                                      | The app helps contact tracing go faster by creating a private digital diary of the places you visit. Users Scan the official QR codes wherever they see them and add manual entries for their visits to other places.                                                                                                                                               | <a href="https://www.health.govt.nz/our-work/diseases-and-conditions/covid-19-novel-coronavirus/covid-19-resources-and-tools/nz-covid-tracer-app">https://www.health.govt.nz/our-work/diseases-and-conditions/covid-19-novel-coronavirus/covid-19-resources-and-tools/nz-covid-tracer-app</a> [123] | <a href="https://github.com/minhealthnz/nz-covid-tracer-app">https://github.com/minhealthnz/nz-covid-tracer-app</a> [125] | [126] [123]<br>[124] [125] |
| VigilantGantry           | This an automated contactless gantry system developed by GovTech's Data Science and Artificial Intelligence Division (DSAID)       | VigilantGantry is an open-source implementation of an AI-driven automated temperature screening gantry that improves the rate of contactless screening by augmenting existing thermal systems. VigilantGantry is excellent for automatically scanning high-traffic sites for symptomatic COVID-19 patients. It helps ground crews keep on the lookout for COVID-19. |                                                                                                                                                                                                                                                                                                     | <a href="https://github.com/dsaigo/vs/vigilantgantry">https://github.com/dsaigo/vs/vigilantgantry</a> [127]               | [128] [127]                |
| lancet-covid-19-database | Developed by Lancet                                                                                                                | The Lancet COVID-19 Database gives users access to the most up-to-date information on COVID-19, such as cases, deaths, recoveries, testing, and other useful indicators for tracking the pandemic's spread and response.                                                                                                                                            |                                                                                                                                                                                                                                                                                                     | <a href="https://github.com/sdsna/lancet-covid-19-database">https://github.com/sdsna/lancet-covid-19-database</a> [130]   | [131] [130]                |
| Covid19Canada            | SDSN developed this application                                                                                                    | This shiny app provides a forecast of Covid-19 cases and Covid-19 information in Canada, including the confirmed new cases of infections, confirmed deaths, and recovery rate. Apart from forecasting, the application utilizes plots to visualize the disease's ten-day forecasts and the accumulated and confirmed data.                                          | <a href="https://artbd.shinyapps.io/covid19canada/">https://artbd.shinyapps.io/covid19canada/</a> [132]                                                                                                                                                                                             | <a href="https://github.com/ccodwg/Covid19Canada">https://github.com/ccodwg/Covid19Canada</a> [133]                       | [3] [132] [133]            |
| COVI-ML                  | This respiratory was developed by the Quebec Artificial Intelligence Institute                                                     | COVI-ML is the Risk model training code for the Covid-19 tracing application. Its repository provides models, infrastructure, and datasets for training deep-learning-based predictors of COVID-19 infectiousness as used in Proactive Contact Tracing.                                                                                                             |                                                                                                                                                                                                                                                                                                     | <a href="https://github.com/milaiqia/COVI-ML">https://github.com/milaiqia/COVI-ML</a> [134]                               | [3][134]                   |

|                              |                                                                                                                                                                                                                                                  |                                                                                                                                                                                                                                                                                        |                                                                                                                                                                                                                                                           |                                                                                                                                   |                        |
|------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------|------------------------|
| Covid-19 model               | Imperial College London developed this application/code.                                                                                                                                                                                         | This code was applied in modeling estimated deaths and infections for COVID-19 from the study "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe "done by Flaxman et al. (2020) [136]                                                                   |                                                                                                                                                                                                                                                           | <a href="https://github.com/ImperialCollegeLondon/covid19model">https://github.com/ImperialCollegeLondon/covid19model</a> [135]   | [136] [135]            |
| The COVID Tracking Project   | Alexis Madrigal created this project through a collaborative volunteer-run effort to track the ongoing COVID-19 pandemic                                                                                                                         | This project collects and publishes data required to understand the COVID-19 outbreak in the United States. All 50 states, five territories, and the District of Columbia participate in the Covid tracking project, which will collect data on COVID-19 testing and patient outcomes. | <a href="https://covidtracking.com/">https://covidtracking.com/</a> [137]                                                                                                                                                                                 | <a href="https://github.com/COVID19Tracking">https://github.com/COVID19Tracking</a> [138]                                         | [139] [137] [138]      |
| Covidmx                      | Covidmx was developed by Federico Garza                                                                                                                                                                                                          | The API was developed to handle Covid-19 open data provided by the Mexican Dirección General de Epidemiología.                                                                                                                                                                         |                                                                                                                                                                                                                                                           | <a href="https://github.com/FedericoGarza/covidmx">https://github.com/FedericoGarza/covidmx</a> [140]                             | [141] [140]            |
| Covid19-Scenarios            | Neherlab developed this tool                                                                                                                                                                                                                     | The Covid-19 Scenarios provide Models of generating trajectories for COVID-19 outbreak and hospital demand. The functioning of this tool is based on the SIR model, which simulates a COVID19 outbreak.                                                                                | <a href="https://covid19-scenarios.org/">https://covid19-scenarios.org/</a> [142]                                                                                                                                                                         | <a href="https://github.com/neherlab/covid19_scenarios">https://github.com/neherlab/covid19_scenarios</a> [143]                   | [3] [142] [143]        |
| covid-chest-imaging-database | This database was developed by NHSX and the British Society of Thoracic Imaging (BSTI). NHSX is a joint unit of National Health Service (NHS) England and the Department of Health and Social Care, supporting local NHS and care organizations. | The database was developed with a respiratory containing tooling related to the NHSX National COVID-19 Chest Image Database (NCCID) to promote research projects in response to the COVID-19 pandemic.                                                                                 |                                                                                                                                                                                                                                                           | <a href="https://github.com/nhsx/covid-chest-imaging-database">https://github.com/nhsx/covid-chest-imaging-database</a> [144]     | [145] [144]            |
| Covid-pass-verifier          | This is an application developed by NHSX                                                                                                                                                                                                         | The COVID Pass Verifier app is the official NHS COVID Pass Verifier for England and Wales. The app is a safe and secure way to check if someone has been appropriately vaccinated against COVID-19, has had a negative test, or has recovered from COVID-19.                           | <a href="https://www.nhs.uk/covid-19-response/nhs-covid-pass-verifier-app/international-covid-pass-verifier-app-user-guide/">https://www.nhs.uk/covid-19-response/nhs-covid-pass-verifier-app/international-covid-pass-verifier-app-user-guide/</a> [146] | <a href="https://github.com/nhsx/covid-pass-verifier">https://github.com/nhsx/covid-pass-verifier</a> [148]                       | [145] [146] [147][148] |
| Covasim                      | The Institute for Disease Modelling developed this simulator                                                                                                                                                                                     | Covasim is a stochastic agent-based simulator for performing COVID-19 analyses.                                                                                                                                                                                                        |                                                                                                                                                                                                                                                           | <a href="https://github.com/InstituteforDiseaseModeling/covasim">https://github.com/InstituteforDiseaseModeling/covasim</a> [149] | [150] [149]            |
| covid-19 Dashboard           | Greg Rafferty developed Covid-19 dashboard                                                                                                                                                                                                       | This is a web dashboard developed to monitor the COVID-19 pandemic. The data used is obtained from Johns Hopkins Center for Systems Science and Engineering.                                                                                                                           | <a href="https://covid-19-raffg.herokuapp.com/">https://covid-19-raffg.herokuapp.com/</a> [151]                                                                                                                                                           | <a href="https://github.com/raffg/covid-19">https://github.com/raffg/covid-19</a> [152]                                           | [3] [152] [151]        |
| Covid-19 R/Python scripts    | Developed by QuKunLa; a Laboratory of Immunogenomics and Precision Medicine, University of Science and Technology of China                                                                                                                       | These are R/Python scripts to analyze single-cell RNA-sequence data from COVID-19 patients.                                                                                                                                                                                            |                                                                                                                                                                                                                                                           | <a href="https://github.com/QuKunLab/COVID-19">https://github.com/QuKunLab/COVID-19</a> [153]                                     | [3] [153]              |
| COVID-19-CT-CXR              | COVID-19-CT-CXR was developed by Peng et al. and Intramural Research Programs of the                                                                                                                                                             | This is a public database of COVID-19 CXR and CT images, which are automatically extracted from COVID-19-relevant articles from the PubMed Central Open Access (PMC-OA)                                                                                                                |                                                                                                                                                                                                                                                           | <a href="https://github.com/ncbi-nlp/COVID-19-CT-CXR">https://github.com/ncbi-nlp/COVID-19-CT-CXR</a> [154]                       | [155] [154]            |

|                                          |                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                    |                                                                                                                                         |                                                                                                                                               |                   |
|------------------------------------------|---------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|-------------------|
|                                          | National Institutes of Health, National Library of Medicine and Clinical Centre.      | Subst.                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                         |                                                                                                                                               |                   |
| covid19-healthsystemcapacity             | This project was developed by the CovidCareMap organization                           | This application assists in better understanding, anticipating, and acting to support and ramp up our health systems' capacity (beds, staffing, ventilators, supplies) to effectively care for a rapidly growing number of active COVID19 patients in need of hospitalization and intensive (ICU) care.                                                                            |                                                                                                                                         | <a href="https://github.com/covidcaremap/covid19-healthsystemcapacity">https://github.com/covidcaremap/covid19-healthsystemcapacity</a> [156] | [3] [156]         |
| CV19 Index                               | The Global Loop team developed this model                                             | The COVID-19 Vulnerability Index (CV19 Index) is a predictive model that identifies persons who are more susceptible to COVID-19 severe problems. The CV19 Index is designed to assist hospitals, federal, state, and local public health agencies, and other healthcare organizations in identifying, planning for, responding to, and reducing COVID-19's impact in their areas. | <a href="https://www.close.dloop.ai/covid-19-index">https://www.close.dloop.ai/covid-19-index</a> [157]                                 | <a href="https://github.com/closedloop-ai/cv19index">https://github.com/closedloop-ai/cv19index</a> [158]                                     | [159] [157] [158] |
| OpenABM-Covid19                          | This model was developed by the Pathogen Dynamics Group of Oxford Big Data Institute. | OpenABM-Covid19 is an agent-based model (ABM) that was created to model the spread of Covid-19 in a city and investigate the impact of passive and active intervention measures.                                                                                                                                                                                                   |                                                                                                                                         | <a href="https://github.com/BDI-pathogens/OpenABM-Covid19">https://github.com/BDI-pathogens/OpenABM-Covid19</a> [160]                         | [3] [160]         |
| COVID-19 vaccination slot booking script | PythonRepo developed this script.                                                     | Is used to automate covid vaccination booking.                                                                                                                                                                                                                                                                                                                                     | <a href="https://pythonrepo.com/repo/pallupz-covid-vaccine-booking">https://pythonrepo.com/repo/pallupz-covid-vaccine-booking</a> [161] | <a href="https://pythonrepo.com/repo/pallupz-covid-vaccine-booking">https://pythonrepo.com/repo/pallupz-covid-vaccine-booking</a> [162]       | [159] [161] [162] |

TABLE II. COVID-19 DATASETS SUMMARY TABLE

| Developer/Industry/Academia/University/Organization Details | Dataset Summary                                                                                                         | Dataset Usage                                                                | Weblink to the Dataset                                                                                                    | References      |
|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------|-----------------|
| Our World in Data                                           | Data on COVID-19 vaccinations that include country-by country statistics of the COVID-19 vaccines administered to date. | Vaccine outreach program.                                                    | <a href="https://ourworldindata.org/covid-vaccinations">https://ourworldindata.org/covid-vaccinations</a> . [164]         | [163][164]      |
|                                                             | Data on COVID-19 confirmed deaths per country.                                                                          | Effects of testing, managing, hospitalization.                               | <a href="https://ourworldindata.org/covid-deaths">https://ourworldindata.org/covid-deaths</a> . [165]                     | [362][165][394] |
|                                                             | Global data of confirmed COVID-19 cases                                                                                 | Effect on travel restriction, intervention programs.                         | <a href="https://ourworldindata.org/covid-cases">https://ourworldindata.org/covid-cases</a> . [166]                       | [393][166]      |
|                                                             | Data on COVID-19 testing, i.e., positivity rate, contact tracing, tests performed per day                               | Pandemic preventive measures                                                 | <a href="https://ourworldindata.org/coronavirus-testing">https://ourworldindata.org/coronavirus-testing</a> . [167]       | [393][167]      |
|                                                             | Data on COVID-19 hospitalization                                                                                        | Monitoring cases to improve impact on available resources.                   | <a href="https://ourworldindata.org/covid-hospitalizations">https://ourworldindata.org/covid-hospitalizations</a> . [168] | [392] [168]     |
|                                                             | COVID-19 mortality risks                                                                                                | Segregation of age groups that may be at risk of dying from the disease.     | <a href="https://ourworldindata.org/mortality-risk-covid">https://ourworldindata.org/mortality-risk-covid</a> . [169]     | [393][169][170] |
|                                                             | Excess mortality due to COVID-19                                                                                        | Segregation of age groups that may be at risk of dying from the disease, and | <a href="https://ourworldindata.org/excess-mortality-covid">https://ourworldindata.org/excess-mortality-covid</a> . [171] | [393][171]      |

|                                                                                    |                                                                                         |                                                                              |                                                                                                                                                                                                                                       |                 |
|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------|
|                                                                                    |                                                                                         | other accelerating factors.                                                  |                                                                                                                                                                                                                                       |                 |
|                                                                                    | Policy responses to the COVID-19 pandemic                                               | Government interventions to curb the spread of the virus.                    | <a href="https://ourworldindata.org/policy-responses-covid">https://ourworldindata.org/policy-responses-covid</a> . [172]                                                                                                             | [393][172][173] |
| The Johns Hopkins University Center for Systems Science and Engineering [JHU CCSE] | COVID-19 Epidemiological Data                                                           | For segmentation of COVID-19 cases based on epidemiological characteristics. | <a href="https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases">https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases</a> . [174]                                                                           | [174]           |
| OCHA                                                                               | COVID-19 number of confirmed cases, deaths, and recoveries by the province in Indonesia | Mobility transmission analysis.                                              | <a href="https://data.humdata.org/dataset/indonesia-covid-19-cases-recoveries-and-deaths-per-province">https://data.humdata.org/dataset/indonesia-covid-19-cases-recoveries-and-deaths-per-province</a> . [175]                       | [175]           |
| World Health Organization                                                          | COVID-19 cases and deaths                                                               | Mobility transmission and mortality analysis.                                | <a href="https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths">https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths</a> . [176]                                                                   | [176]           |
| Blavatnik School of Government, University of Oxford                               | OXFORD COVID-19 Government Response Stringency index                                    | Government measures                                                          | <a href="https://data.humdata.org/dataset/oxford-covid-19-government-response-tracker">https://data.humdata.org/dataset/oxford-covid-19-government-response-tracker</a> . [177]                                                       | [177]           |
| HDX                                                                                | COVID-19 Vaccinations                                                                   | Rate of vaccine drives                                                       | <a href="https://data.humdata.org/dataset/covid-19-vaccinations">https://data.humdata.org/dataset/covid-19-vaccinations</a> . [178]                                                                                                   | [178]           |
| World Food Program                                                                 | COVID-19 global airline information and travel restriction                              | Global Monitoring.                                                           | <a href="https://data.humdata.org/dataset/covid-19-global-travel-restrictions-and-airline-information">https://data.humdata.org/dataset/covid-19-global-travel-restrictions-and-airline-information</a> . [179]                       | [179]           |
| HDX                                                                                | COVID-19 cases and deaths in the United States                                          | Reporting cases at a national level.                                         | <a href="https://data.humdata.org/dataset/nyt-covid-19-data">https://data.humdata.org/dataset/nyt-covid-19-data</a> . [180]                                                                                                           | [180]           |
| HDX                                                                                | Total number of COVID-19 tests performed per country                                    | Monitoring cases                                                             | <a href="https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country">https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country</a> . [181]                                                           | [181]           |
| UNESCO                                                                             | Global school closures                                                                  | Area segmentation                                                            | <a href="https://data.humdata.org/dataset/global-school-closures-covid19">https://data.humdata.org/dataset/global-school-closures-covid19</a> . [182]                                                                                 | [182]           |
| Meta                                                                               | FAIR COVID-19 US County Forecast                                                        | Country-level forecast.                                                      | <a href="https://data.humdata.org/dataset/fair-covid-dataset">https://data.humdata.org/dataset/fair-covid-dataset</a> . [183]                                                                                                         | [183]           |
| CARE Bangladesh                                                                    | District Wise Quarantine for COVID-19                                                   | Reporting cases at a national level.                                         | <a href="https://data.humdata.org/dataset/district-wise-quarantine-for-covid-19">https://data.humdata.org/dataset/district-wise-quarantine-for-covid-19</a> . [184]                                                                   | [184]           |
| HDX                                                                                | COVID-19 Impact on Humanitarian Operations Data Viz inputs                              | Reporting humanitarian activities at a national level.                       | <a href="https://data.humdata.org/dataset/covid-19-data-visual-inputs">https://data.humdata.org/dataset/covid-19-data-visual-inputs</a> . [185]                                                                                       | [185]           |
| OCHA Venezuela                                                                     | COVID-19 sub-national data                                                              | Reporting cases at a national level.                                         | <a href="https://data.humdata.org/dataset/corona-virus-covid-19-cases-and-deaths-in-venezuela">https://data.humdata.org/dataset/corona-virus-covid-19-cases-and-deaths-in-venezuela</a> . [186]                                       | [186]           |
| OCHA FISS                                                                          | Global Humanitarian Operational Presence Who, What, Where [3W] Portal                   | Reporting humanitarian activities at a global level.                         | <a href="https://data.humdata.org/dataset/ocha-global-humanitarian-operational-presence-who-what-where-3w-portal">https://data.humdata.org/dataset/ocha-global-humanitarian-operational-presence-who-what-where-3w-portal</a> . [187] | [187]           |
| ACAPS                                                                              | COVID-19 Government Measures Dataset                                                    | Reporting government measures at a global level.                             | <a href="https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset">https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset</a> . [188]                                                           | [188]           |
| HDX                                                                                | Europe COVID-19 subnational cases                                                       | COVID-19 infected area segmentation.                                         | <a href="https://data.humdata.org/dataset/europe-covid-19-subnational-cases">https://data.humdata.org/dataset/europe-covid-19-subnational-cases</a> . [190]                                                                           | [190]           |
| OCHA Philippines                                                                   | Philippines COVID-19 response.                                                          | Reporting government measures at a national level.                           | <a href="https://data.humdata.org/dataset/philippines-covid-19-response-who-does-what-where">https://data.humdata.org/dataset/philippines-covid-19-response-who-does-what-where</a> . [191]                                           | [191]           |

|                                               |                                                                          |                                                                                                                                      |                                                                                                                                                                                                                       |       |
|-----------------------------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
| Code for Venezuela                            | COVID-19 education impact survey                                         | Monitoring impact on a national level                                                                                                | <a href="https://data.humdata.org/dataset/open_one_time_covid_education_impact">https://data.humdata.org/dataset/open_one_time_covid_education_impact</a> . [192]                                                     | [192] |
| iMMAP                                         | Google mobility report                                                   | Mobility transmission analysis.                                                                                                      | <a href="https://data.humdata.org/dataset/google-mobility-report">https://data.humdata.org/dataset/google-mobility-report</a> . [193]                                                                                 | [193] |
| Humanitarian Emergency Report Africa [HERA]   | Subnational data on Covid 19 cases per day                               | COVID-19 infected area segmentation.                                                                                                 | <a href="https://data.humdata.org/dataset/nigeria_covid19_subnational">https://data.humdata.org/dataset/nigeria_covid19_subnational</a> . [194]                                                                       | [194] |
| HDX                                           | Worldwide geographic distribution of COVID-19 cases                      | COVID-19 infected area segmentation.                                                                                                 | <a href="https://data.humdata.org/dataset/ecdc-covid-19">https://data.humdata.org/dataset/ecdc-covid-19</a> . [195]                                                                                                   | [195] |
| World Health Organization                     | Immunization campaigns impacted due to COVID-19.                         | Mobility transmission analysis                                                                                                       | <a href="https://data.humdata.org/dataset/immunization-campaigns-impacted">https://data.humdata.org/dataset/immunization-campaigns-impacted</a> . [196]                                                               | [196] |
| HDX                                           | Excess mortality during COVID-19 pandemic                                | Segregation of age groups that may be at risk of dying from the disease.                                                             | <a href="https://data.humdata.org/dataset/financial-times-excess-mortality-during-covid-19-pandemic-data">https://data.humdata.org/dataset/financial-times-excess-mortality-during-covid-19-pandemic-data</a> . [197] | [197] |
| HDX                                           | COVID-19 subnational cases in Palestine                                  | Reporting cases at a national level.                                                                                                 | <a href="https://data.humdata.org/dataset/state-of-palestine-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/state-of-palestine-coronavirus-covid-19-subnational-cases</a> . [198]           | [198] |
| Meta                                          | Impact survey and trends on COVID-19                                     | Reporting cases at a national level.                                                                                                 | <a href="https://data.humdata.org/dataset/covid-19-symptom-map">https://data.humdata.org/dataset/covid-19-symptom-map</a> . [199]                                                                                     | [199] |
| HDX                                           | COVID-19 vaccine doses are given to humanitarian resource plan countries | Forecasts on dose availability and actual deliveries                                                                                 | <a href="https://data.humdata.org/dataset/covid-19-vaccine-doses-in-hrp-countries">https://data.humdata.org/dataset/covid-19-vaccine-doses-in-hrp-countries</a> . [200]                                               | [200] |
| World Bank Group                              | World Bank indicators of interest to the COVID-19 outbreak               | Data for use in response, modeling analysis                                                                                          | <a href="https://data.humdata.org/dataset/world-bank-indicators-of-interest-to-the-covid-19-outbreak">https://data.humdata.org/dataset/world-bank-indicators-of-interest-to-the-covid-19-outbreak</a> . [201]         | [201] |
| Global Health 50/50                           | Gender and COVID-19 project                                              | Exploring how gender may be driving the higher proportion of reported deaths in men among confirmed cases so far.                    | <a href="http://globalhealth5050.org/covid19">http://globalhealth5050.org/covid19</a> [202]                                                                                                                           | [202] |
| World Bank Group                              | Harmonized data on Household COVID-19 monitoring surveys                 | Data analysis and trend checking                                                                                                     | <a href="https://data.humdata.org/dataset/harmonized-covid-19-household-monitoring-surveys">https://data.humdata.org/dataset/harmonized-covid-19-household-monitoring-surveys</a> [203]                               | [203] |
| Humanitarian Emergency Response Africa [HERA] | African continent Covid 19 cases                                         | Data analysis and trendsetting                                                                                                       | <a href="https://data.humdata.org/dataset/covid19_africa_contenital_infections-recoveries-deaths">https://data.humdata.org/dataset/covid19_africa_contenital_infections-recoveries-deaths</a> [204]                   | [204] |
| Dalberg                                       | Developing countries' government action on COVID-19                      | non-pharmaceutical interventions                                                                                                     | <a href="https://data.humdata.org/dataset/government-actions-on-covid-19">https://data.humdata.org/dataset/government-actions-on-covid-19</a> [205]                                                                   | [205] |
| Meta                                          | Survey on preventative health                                            | Monitor and understand people's knowledge and practices about COVID-19 to improve communications and their response to the pandemic. | <a href="https://data.humdata.org/dataset/preventive-health-survey">https://data.humdata.org/dataset/preventive-health-survey</a> [206]                                                                               | [206] |
| International Organization for Migration      | Information on populations within the Far North region of Cameroon       | Providing regular, accurate, and updated data to better support the response of the Government of Cameroon and the humanitarian      | <a href="https://data.humdata.org/dataset/cameroon-baseline-assessment-data-iom-dtm">https://data.humdata.org/dataset/cameroon-baseline-assessment-data-iom-dtm</a> [207]                                             | [207] |

|                                        |                                                                        |                                                                                    |                                                                                                                                                                                                                                                                                       |       |
|----------------------------------------|------------------------------------------------------------------------|------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
|                                        |                                                                        | community.                                                                         |                                                                                                                                                                                                                                                                                       |       |
| HDX                                    | Covax round 6 allocations                                              | Monitoring of Covax vaccine allocations                                            | <a href="https://data.humdata.org/dataset/covax-round-6-allocations">https://data.humdata.org/dataset/covax-round-6-allocations</a> [208]                                                                                                                                             | [208] |
| Humanitarian Emergency Response Africa | COVID-19 subnational data in Burkina Faso                              | Reporting Covid data at National level                                             | <a href="https://data.humdata.org/dataset/burkinafaso_covid19_subnational">https://data.humdata.org/dataset/burkinafaso_covid19_subnational</a> [209]                                                                                                                                 | [209] |
| Metabiota                              | Spatiotemporal data for COVID-19 deaths and cases.                     | Data analysis and monitoring                                                       | <a href="https://data.humdata.org/dataset/2019-novel-coronavirus-cases">https://data.humdata.org/dataset/2019-novel-coronavirus-cases</a> [210]                                                                                                                                       | [210] |
| HDX                                    | COVID-19 subnational data for Afghanistan                              | Data analysis and reporting on a national level                                    | <a href="https://data.humdata.org/dataset/afghanistan-covid-19-statistics-per-province">https://data.humdata.org/dataset/afghanistan-covid-19-statistics-per-province</a> [211]                                                                                                       | [211] |
| Cuebiq Inc                             | COVID-19 mobility data for Italy                                       | Monitoring mobility changes in Italy since lockdown                                | <a href="https://data.humdata.org/dataset/covid-19-mobility-italy">https://data.humdata.org/dataset/covid-19-mobility-italy</a> [212]                                                                                                                                                 | [212] |
| Humanitarian Emergency Response Africa | COVID-19 subnational cases in Africa                                   | Reporting COVID-19 cases on a national level                                       | <a href="https://data.humdata.org/dataset/africa-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/africa-coronavirus-covid-19-subnational-cases</a> [213]                                                                                                     | [213] |
| Qatar Computing Research Institute     | Twitter data geographic distribution of COVID-19                       | Geographical distribution of twitter users and tweets regarding COVID-19 pandemic. | <a href="https://data.humdata.org/dataset/covid-19-twitter-data-geographic-distribution">https://data.humdata.org/dataset/covid-19-twitter-data-geographic-distribution</a> [214]                                                                                                     | [214] |
| ACAPS                                  | Secondary impacts of Covid 19 on a global scale                        | Aid Decision-making on addressing wider effects of COVID-19                        | <a href="https://data.humdata.org/dataset/global-covid-19-secondary-impacts">https://data.humdata.org/dataset/global-covid-19-secondary-impacts</a> [215]                                                                                                                             | [215] |
| Humanitarian Emergency Response Africa | COVID-19 city level in Burkina Faso                                    | Reporting Covid data at a city level                                               | <a href="https://data.humdata.org/dataset/burkinafaso_covid19_city-level">https://data.humdata.org/dataset/burkinafaso_covid19_city-level</a> [216]                                                                                                                                   | [216] |
| Hub Latin America                      | The COVID-19 mortality rate in Lima, Peru                              | Reporting, analysis of COVID-19 death rates in Lima                                | <a href="https://data.humdata.org/dataset/peru-covid19-mortality-rate-in-lima">https://data.humdata.org/dataset/peru-covid19-mortality-rate-in-lima</a> [217]                                                                                                                         | [217] |
| Infoculture                            | COVID-19 cases in Moscow                                               | Statistics                                                                         | <a href="https://data.humdata.org/dataset/covid-19-cases-data-in-moscow">https://data.humdata.org/dataset/covid-19-cases-data-in-moscow</a> [218]                                                                                                                                     | [218] |
| HDX                                    | Social measures and public health applied during COVID-19              | Analysis and reporting.                                                            | <a href="https://data.humdata.org/dataset/world-global-database-of-public-health-and-social-measures-applied-during-the-covid-19-pandemic">https://data.humdata.org/dataset/world-global-database-of-public-health-and-social-measures-applied-during-the-covid-19-pandemic</a> [219] | [219] |
| Mobile Accord, Inc [GeoPoll]           | Impact and perceptions of Coronavirus in Sub-Saharan African countries | Analysis and reporting                                                             | <a href="https://data.humdata.org/dataset/covid-19-impacts-africa">https://data.humdata.org/dataset/covid-19-impacts-africa</a> [220]                                                                                                                                                 | [220] |
| HDX                                    | Subnational COVID-19 cases for Iraq                                    | Reporting Covid data at National level                                             | <a href="https://data.humdata.org/dataset/iraq-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/iraq-coronavirus-covid-19-subnational-cases</a> [221]                                                                                                         | [221] |
| HDX                                    | Covid 19 related funding from IATI                                     | Monitoring of funding use in fighting COVID-19                                     | <a href="https://data.humdata.org/dataset/iati-covid19-funding">https://data.humdata.org/dataset/iati-covid19-funding</a> [222]                                                                                                                                                       | [222] |
| HDX                                    | Gavi and World Bank COVID-19 vaccine funding                           | Fund disbursement and support for COVID-19                                         | <a href="https://data.humdata.org/dataset/world-bank-and-gavi-vaccine-financing">https://data.humdata.org/dataset/world-bank-and-gavi-vaccine-financing</a> [223]                                                                                                                     | [223] |
| Code for Venezuela                     | Survey on COVID-19 impact                                              | Data analysis and interpretation                                                   | <a href="https://data.humdata.org/dataset/open_one_time_covid_impact">https://data.humdata.org/dataset/open_one_time_covid_impact</a> [224]                                                                                                                                           | [224] |
| Humanitarian Emergency Response Africa | COVID-19 cases in Ethiopia                                             | Reporting cases at a national level.                                               | <a href="https://data.humdata.org/dataset/ethiopia-covid19-cases">https://data.humdata.org/dataset/ethiopia-covid19-cases</a> [225]                                                                                                                                                   | [225] |
| World Bank Group                       | High frequency indicators for COVID-19                                 | Data analysis and interpretation                                                   | <a href="https://data.humdata.org/dataset/covid-19-high-frequency-indicators">https://data.humdata.org/dataset/covid-19-high-frequency-indicators</a> [226]                                                                                                                           | [226] |

|                                                |                                                                                                |                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                     |            |
|------------------------------------------------|------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| OCHA FISS                                      | Global humanitarian response plan COVID-19 administrative boundaries and population-statistics | Reporting cases at a national level.                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/global-humanitarian-response-plan-covid-19-administrative-boundaries-and-population-statistics">https://data.humdata.org/dataset/global-humanitarian-response-plan-covid-19-administrative-boundaries-and-population-statistics</a> [227] | [227]      |
| INFORM                                         | Inform Risk Index for COVID-19, Version 0.1.4                                                  | Support prioritization of preparedness and early response actions for the direct impacts of the pandemic and identify countries where secondary effects are likely to have the most critical humanitarian consequences. | <a href="https://data.humdata.org/dataset/inform-covid-19-risk-index-version-0-1-4">https://data.humdata.org/dataset/inform-covid-19-risk-index-version-0-1-4</a> [228]                                                                                                             | [228]      |
| Safeture                                       | COVID-19 subnational cases in Kazakhstan                                                       | For data analysis and interpretation                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/kazakhstan-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/kazakhstan-coronavirus-covid-19-subnational-cases</a> . [229]                                                                                         | [229]      |
| OCHA Philippines                               | COVID-19-operational presence risk communication and community engagement in the Philippines   | Risk communication and community engagement                                                                                                                                                                             | <a href="https://data.humdata.org/dataset/philippines-covid-19-operational-presence-risk-communication-and-community-engagement-rcce">https://data.humdata.org/dataset/philippines-covid-19-operational-presence-risk-communication-and-community-engagement-rcce</a> . [230]       | [230]      |
| Hub Latin America                              | Epidemiological and hospital indicators on COVID-19 in Ouro Preto, Brazil                      | For data analysis and interpretation                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/brazil-epidemiological-and-hospital-indicators-on-covid-19-in-ouro-preto">https://data.humdata.org/dataset/brazil-epidemiological-and-hospital-indicators-on-covid-19-in-ouro-preto</a> . [231]                                           | [231]      |
| Safeture                                       | COVID-19 subnational cases in Oman                                                             | Reporting cases at a national level.                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/oman-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/oman-coronavirus-covid-19-subnational-cases</a> . [232]                                                                                                     | [232]      |
| Humanitarian Emergency Response Africa         | Coronavirus [COVID-19] City level cases for Mauritania                                         | Reporting cases at a national level.                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/mauritania-coronavirus-covid-19-city-level">https://data.humdata.org/dataset/mauritania-coronavirus-covid-19-city-level</a> . [233]                                                                                                       | [233]      |
| UNICEF Data and Analytics [HQ]                 | Tracking children's situation during COVID-19                                                  | Data analysis and interpretation                                                                                                                                                                                        | <a href="https://data.humdata.org/dataset/rapid-situation-tracking-for-covid-19-socioeconomic-impacts">https://data.humdata.org/dataset/rapid-situation-tracking-for-covid-19-socioeconomic-impacts</a> . [234]                                                                     | [234]      |
| Humanitarian Emergency Response Africa         | COVID-19 recoveries in Africa on a national level                                              | Data analysis and interpretation                                                                                                                                                                                        | <a href="https://data.humdata.org/dataset/africa-covid-19-recovered-cases">https://data.humdata.org/dataset/africa-covid-19-recovered-cases</a> . [235]                                                                                                                             | [235]      |
| Mobile Accord, Inc. [GeoPoll]                  | COVID-19 vaccines and impacts accepted in Sub-Saharan Africa                                   | Data analysis and interpretation                                                                                                                                                                                        | <a href="https://data.humdata.org/dataset/covid19-impacts-and-vaccine-acceptance-in-sub-saharan-africa">https://data.humdata.org/dataset/covid19-impacts-and-vaccine-acceptance-in-sub-saharan-africa</a> . [236]                                                                   | [236]      |
| United Nations Development Coordination Office | UN Collective Results on the COVID-19 Socioeconomic Response in 2020                           | Monitor the progress and achievements of UNCT's collective actions in socio-economic response.                                                                                                                          | <a href="https://data.humdata.org/dataset/un-collective-results-on-the-covid-19-socioeconomic-response-in-2020">https://data.humdata.org/dataset/un-collective-results-on-the-covid-19-socioeconomic-response-in-2020</a> . [237]                                                   | [237]      |
| Mobile Accord, Inc. [GeoPoll]                  | Economic impact of COVID-19 in Sub Saharan Africa                                              | Data interpretation and analysis                                                                                                                                                                                        | <a href="https://data.humdata.org/dataset/economic-impact-of-covid-19-in-sub-saharan-africa">https://data.humdata.org/dataset/economic-impact-of-covid-19-in-sub-saharan-africa</a> . [238]                                                                                         | [238]      |
| HDX                                            | COVID-19 subnational cases in Myanmar                                                          | Reporting cases at a national level.                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/myanmar-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/myanmar-coronavirus-covid-19-subnational-cases</a> . [239]                                                                                               | [239]      |
| Safeture                                       | COVID-19 sub-national cases in Ghana                                                           | Reporting cases at a national level.                                                                                                                                                                                    | <a href="https://data.humdata.org/dataset/ghana-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/ghana-coronavirus-covid-19-subnational-cases</a> [240]                                                                                                     | [240]      |
| Insecurity Insight                             | Covid 19 and Aid Security                                                                      | To help aid agencies meet the duty of care obligations to staff and reach people in need.                                                                                                                               | <a href="https://data.humdata.org/dataset/aid-security-and-covid-19">https://data.humdata.org/dataset/aid-security-and-covid-19</a> . [241]                                                                                                                                         | [241]      |
| Infoculture                                    | Registry of Russian NGO's affected by COVID-19                                                 | For data analysis and interpretation.                                                                                                                                                                                   | <a href="https://data.humdata.org/dataset/ngos-affected-by-covid19-russia">https://data.humdata.org/dataset/ngos-affected-by-covid19-russia</a> . [242]                                                                                                                             | [242][243] |

|                                          |                                                                      |                                                     |                                                                                                                                                                                                                                     |            |
|------------------------------------------|----------------------------------------------------------------------|-----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
|                                          |                                                                      |                                                     |                                                                                                                                                                                                                                     |            |
| HDX                                      | Facility Interim Distribution Forecast for Covax                     | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/covax-facility-interim-distribution-forecast">https://data.humdata.org/dataset/covax-facility-interim-distribution-forecast</a> . [244]                                                   | [244]      |
| UNHCR - The UN Refugee Agency            | Socio-economic impact of COVID-19 on refugees in Kenya               | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-socioeconomic-impact-of-covid-19-on-pocs-in-kenya-v2-2">https://data.humdata.org/dataset/unhcr-ken-2020-socioeconomic-impact-of-covid-19-on-pocs-in-kenya-v2-2</a> . [245] | [245]      |
| Humanitarian Emergency Response Africa   | COVID-19 city level cases in Togo                                    | Reporting cases at a national level.                | <a href="https://data.humdata.org/dataset/togo-coronavirus-covid-19-city-level">https://data.humdata.org/dataset/togo-coronavirus-covid-19-city-level</a> . [246]                                                                   | [246]      |
| UNICEF Data and Analytics                | Indicators of interest to COVID-19 data at UNICEF                    | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/unicef-indicators-of-interest-to-the-covid-19-outbreak">https://data.humdata.org/dataset/unicef-indicators-of-interest-to-the-covid-19-outbreak</a> . [247]                               | [247]      |
| HDX                                      | COVID-19 subnational cases in Mozambique                             | Reporting cases at a national level.                | <a href="https://data.humdata.org/dataset/mozambique-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/mozambique-coronavirus-covid-19-subnational-cases</a> . [248]                                         | [248]      |
| HDX                                      | COVID-19 subnational cases for Haiti                                 | Reporting cases at a national level.                | <a href="https://data.humdata.org/dataset/haiti-covid-19-subnational-cases">https://data.humdata.org/dataset/haiti-covid-19-subnational-cases</a> [249]                                                                             | [249]      |
| OCHA HQ                                  | COVID-19 Pandemic induced Humanitarian Access Constraints            | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/constraints-faced-by-people-due-to-covid-19-outbreak">https://data.humdata.org/dataset/constraints-faced-by-people-due-to-covid-19-outbreak</a> . [250]                                   | [250]      |
| OCHA Philippines                         | 2020 Significant events happening in Philippines                     | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/philippines-2020-significant-events">https://data.humdata.org/dataset/philippines-2020-significant-events</a> . [251]                                                                     | [251]      |
| UNHCR - The UN Refugee Agency            | Socio-economic impact of COVID-19 on refugees in Kenya round 5       | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-covid-round5-v2-1">https://data.humdata.org/dataset/unhcr-ken-2020-covid-round5-v2-1</a> . [252]                                                                           | [252][253] |
| OCHA Sudan                               | COVID-19 response and preparedness 4W in Sudan                       | COVID-19 response outcomes                          | <a href="https://data.humdata.org/dataset/sudan-covid-19-preparedness-and-response-4w">https://data.humdata.org/dataset/sudan-covid-19-preparedness-and-response-4w</a> . [254]                                                     | [254]      |
| Indonesian Red Cross [PMI]               | Community Feedback by Indonesian Red Cross [PMI]                     | COVID-19 response outcomes                          | <a href="https://data.humdata.org/dataset/community-feedback-by-indonesian-red-cross-pmi">https://data.humdata.org/dataset/community-feedback-by-indonesian-red-cross-pmi</a> . [255]                                               | [255]      |
| Johns Hopkins Applied Physics Lab        | Projected COVID-19 subnational cases in Sudan                        | For data analysis and interpretation                | <a href="https://data.humdata.org/dataset/sudan-projected-covid-19-sub-national-cases">https://data.humdata.org/dataset/sudan-projected-covid-19-sub-national-cases</a> . [256]                                                     | [256]      |
| International Organization for Migration | IATA travel restriction monitoring                                   | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/travel-restriction-monitoring-iata-covid-19-iom-dtm">https://data.humdata.org/dataset/travel-restriction-monitoring-iata-covid-19-iom-dtm</a> . [257]                                     | [257]      |
| OCHA ROWCA                               | COVID-19 situation in West and Central Africa                        | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/west-and-central-africa-coronavirus-covid-19-situation">https://data.humdata.org/dataset/west-and-central-africa-coronavirus-covid-19-situation</a> . [258]                               | [258]      |
| Johns Hopkins Applied Physics Lab        | Projected COVID-19 subnational cases for Somalia                     | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/somalia-projected-covid-19-sub-national-cases">https://data.humdata.org/dataset/somalia-projected-covid-19-sub-national-cases</a> [259]                                                   | [259]      |
| OCHA Ethiopia                            | COVID-19 sub-national cases for Ethiopia.                            | Reporting cases at a national level.                | <a href="https://data.humdata.org/dataset/ethiopia-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/ethiopia-coronavirus-covid-19-subnational-cases</a> [260]                                               | [260]      |
| Infoculture                              | COVID-19 cases in Russia                                             | Reporting cases at a national level.                | <a href="https://data.humdata.org/dataset/covid-19-cases-data-in-russia">https://data.humdata.org/dataset/covid-19-cases-data-in-russia</a> . [261]                                                                                 | [261]      |
| Mobile Accord, Inc. [GeoPoll]            | Ongoing impacts of COVID-19 in Sub-Saharan Africa                    | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/ongoing-impacts-of-covid-19-in-sub-saharan-africa">https://data.humdata.org/dataset/ongoing-impacts-of-covid-19-in-sub-saharan-africa</a> . [262]                                         | [262]      |
| OCHA HQ                                  | Global appeals and plans of COVID-19 around the globe                | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/covid-19-global-appeals-and-plans">https://data.humdata.org/dataset/covid-19-global-appeals-and-plans</a> . [263]                                                                         | [263]      |
| Mobile Accord, Inc. [GeoPoll]            | Community perception and knowledge of Covid 19 in sub-Saharan Africa | For data analysis and interpretation.               | <a href="https://data.humdata.org/dataset/coronavirus-in-sub-saharan-africa">https://data.humdata.org/dataset/coronavirus-in-sub-saharan-africa</a> [264]                                                                           | [264]      |
| OCHA HQ                                  | COVID-19 allocations for CERF and CBPF                               | Monitoring resource allocation                      | <a href="https://data.humdata.org/dataset/cerf-covid-19-allocations">https://data.humdata.org/dataset/cerf-covid-19-allocations</a> . [265]                                                                                         | [265]      |
| INFORM                                   | INFORM COVID-19 comparability and analysis tool                      | Identification of countries at risk from health and | <a href="https://data.humdata.org/dataset/inform-covid-analysis-v01">https://data.humdata.org/dataset/inform-covid-analysis-v01</a> . [266]                                                                                         | [266]      |

|                                          |                                                                                     |                                                                                                                                                                |                                                                                                                                                                                                                                                         |            |
|------------------------------------------|-------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
|                                          |                                                                                     | humanitarian impacts of COVID-19 that could overwhelm current national response capacity, and therefore lead to a need for additional international assistance |                                                                                                                                                                                                                                                         |            |
| OCHA HQ                                  | Covid 19 impacts, mitigation, and humanitarian access constraint.                   | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/covid19-humanitarian-access">https://data.humdata.org/dataset/covid19-humanitarian-access</a> . [267]                                                                                                         | [267]      |
| HDX                                      | LSHTM COVID-19 Projections.                                                         | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/lshtm-covid-19-projections">https://data.humdata.org/dataset/lshtm-covid-19-projections</a> [268]                                                                                                             | [268]      |
| UNICEF ESARO                             | UNICEF COVID-19 response and situation in Eastern and Southern Africa               | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/eastern-and-southern-africa-covid-19-unicef-situation-and-response">https://data.humdata.org/dataset/eastern-and-southern-africa-covid-19-unicef-situation-and-response</a> [269]                             | [269]      |
| OCHA Mali                                | COVID-19 subnational cases in Mali                                                  | Reporting cases at a national level.                                                                                                                           | <a href="https://data.humdata.org/dataset/mali-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/mali-coronavirus-covid-19-subnational-cases</a> . [270]                                                                         | [270]      |
| OCHA HQ                                  | Economic exposure index for COVID-19                                                | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/covid-19-economic-exposure-index">https://data.humdata.org/dataset/covid-19-economic-exposure-index</a> . [271]                                                                                               | [271]      |
| OCHA Somalia                             | COVID-19 sub-national cases for Somalia                                             | Reporting cases at a national level.                                                                                                                           | <a href="https://data.humdata.org/dataset/somalia-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/somalia-coronavirus-covid-19-subnational-cases</a> . [272]                                                                   | [272][273] |
| Johns Hopkins Applied Physics Lab        | Projected COVID-19 subnational cases for Afghanistan                                | Reporting cases at a national level.                                                                                                                           | <a href="https://data.humdata.org/dataset/afghanistan-projected-covid-19-sub-national-cases">https://data.humdata.org/dataset/afghanistan-projected-covid-19-sub-national-cases</a> . [275]                                                             | [274][275] |
| UNHCR - The UN Refugee Agency            | Testing, knowledge, and mask-wearing                                                | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/unhcr-bgd-2020-covid-mwtk-v2-1">https://data.humdata.org/dataset/unhcr-bgd-2020-covid-mwtk-v2-1</a> . [276]                                                                                                   | [276]      |
| Uganda Red Cross Society                 | COVID-19 risk index                                                                 | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/covid19_risk_index-zip">https://data.humdata.org/dataset/covid19_risk_index-zip</a> . [277]                                                                                                                   | [277]      |
| HDX                                      | COVID-19 subnational cases for the Democratic Republic of Congo                     | Reporting cases at a national level.                                                                                                                           | <a href="https://data.humdata.org/dataset/democratic-republic-of-the-congo-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/democratic-republic-of-the-congo-coronavirus-covid-19-subnational-cases</a> . [278]                 | [278]      |
| HDX                                      | COVID-19 subnational data for Libya                                                 | Reporting cases at a national level.                                                                                                                           | <a href="https://data.humdata.org/dataset/libya-coronavirus-covid-19-subnational-cases">https://data.humdata.org/dataset/libya-coronavirus-covid-19-subnational-cases</a> . [279]                                                                       | [279]      |
| UNHCR - The UN Refugee Agency            | Round 2 Socio-economic impacts of COVID-19 on refugees in Kenya                     | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-socioeconomic-impact-of-covid-19-on-pocs-in-kenya-round2-v1-0">https://data.humdata.org/dataset/unhcr-ken-2020-socioeconomic-impact-of-covid-19-on-pocs-in-kenya-round2-v1-0</a> . [280]       | [280]      |
| International Organization for Migration | Cameroon COVID-19 Mobility Restriction - Point of Entries - [IOM DTM]               | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/cameroon-covid-19-mobility-restriction-point-of-entries-iom-dtm">https://data.humdata.org/dataset/cameroon-covid-19-mobility-restriction-point-of-entries-iom-dtm</a> . [281]                                 | [281]      |
| UNHCR - The UN Refugee Agency            | A panel study of the socio-economic impacts of COVID-19 on refugees living in Kenya | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-covid-panel-v2-1">https://data.humdata.org/dataset/unhcr-ken-2020-covid-panel-v2-1</a> . [282]                                                                                                 | [282]      |
| Johns Hopkins Applied Physics Lab        | Projected COVID-19 subnational cases for Iraq                                       | Reporting cases at a national level.                                                                                                                           | <a href="https://data.humdata.org/dataset/iraq-projected-covid-19-sub-national-cases">https://data.humdata.org/dataset/iraq-projected-covid-19-sub-national-cases</a> . [282]                                                                           | [283]      |
| UNHCR - The UN Refugee Agency            | Assessment of COVID-19 socio-economic impacts on Persons of concern to UNHCR        | Reporting cases at national level.                                                                                                                             | <a href="https://data.humdata.org/dataset/unhcr-nga-2020-sea-covid19-v2-1">https://data.humdata.org/dataset/unhcr-nga-2020-sea-covid19-v2-1</a> . [284]                                                                                                 | [284][285] |
| UNHCR - The UN Refugee Agency            | Assessment of COVID-19 impact on livelihoods of refugees in Zambia                  | For data analysis and interpretation.                                                                                                                          | <a href="https://data.humdata.org/dataset/ddi-zam-unhcr-covid19-impact-assessment-on-refugee-livelihoods-zambia-july-2020">https://data.humdata.org/dataset/ddi-zam-unhcr-covid19-impact-assessment-on-refugee-livelihoods-zambia-july-2020</a> . [286] | [286]      |

|                                                    |                                                                                                                                                             |                                                                                              |                                                                                                                                                                                                                                                                                   |            |
|----------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Hub Latin America                                  | symptomatology related to the coronavirus COVID-19 in Ecuador                                                                                               | Data exploration, analysis                                                                   | <a href="https://data.humdata.org/dataset/symptomatology-ecu911-santa-cruz-monthly-2018-2021">https://data.humdata.org/dataset/symptomatology-ecu911-santa-cruz-monthly-2018-2021</a> . [287]                                                                                     | [287][288] |
| HDX                                                | Health facilities by province in Afghanistan                                                                                                                | For data analysis and interpretation.                                                        | <a href="https://data.humdata.org/dataset/afghanistan-covid-19-health-facilities-by-province">https://data.humdata.org/dataset/afghanistan-covid-19-health-facilities-by-province</a> . [289]                                                                                     | [289]      |
| ACAPS                                              | COVID-19 humanitarian exceptions                                                                                                                            | For data analysis and interpretation.                                                        | <a href="https://data.humdata.org/dataset/acaps-covid-19-humanitarian-exemptions-dataset">https://data.humdata.org/dataset/acaps-covid-19-humanitarian-exemptions-dataset</a> [290]                                                                                               | [290]      |
| International Organization for Migration           | COVID-19 mobility and preparedness updates in South Sudan.                                                                                                  | For data analysis and interpretation.                                                        | <a href="https://data.humdata.org/dataset/south-sudan-covid-19-mobility-and-preparedness-updates-iom-dtm">https://data.humdata.org/dataset/south-sudan-covid-19-mobility-and-preparedness-updates-iom-dtm</a> . [291]                                                             | [291]      |
| UNHCR - The UN Refugee Agency                      | Socio-economic impacts of COVID-19 on refugees living in Kenya, Round 1                                                                                     | For data analysis and interpretation.                                                        | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-covid-round1-v2-2">https://data.humdata.org/dataset/unhcr-ken-2020-covid-round1-v2-2</a> . [292]                                                                                                                         | [292]      |
| UNHCR - The UN Refugee Agency                      | Socio-economic impacts of COVID-19 on refugees living in Kenya, Round 4                                                                                     | For data analysis and interpretation.                                                        | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-covid-round4-v2-1">https://data.humdata.org/dataset/unhcr-ken-2020-covid-round4-v2-1</a> . [293]                                                                                                                         | [293]      |
| UNHCR - The UN Refugee Agency                      | Socio-economic impacts of COVID-19 on refugees living in Kenya, Round 3                                                                                     | For data analysis and interpretation.                                                        | <a href="https://data.humdata.org/dataset/unhcr-ken-2020-covid-round3-v2-1">https://data.humdata.org/dataset/unhcr-ken-2020-covid-round3-v2-1</a> . [294]                                                                                                                         | [294]      |
| European Centre for Disease Prevention and Control | COVID-19 vaccination in the EU/EEA                                                                                                                          | Vaccine administration updates                                                               | <a href="https://www.ecdc.europa.eu/en/publications-data/data-COVID-19-vaccination-eu-eea">https://www.ecdc.europa.eu/en/publications-data/data-COVID-19-vaccination-eu-eea</a> . [295]                                                                                           | [295]      |
|                                                    | Data on the daily number of new reported COVID-19 cases and deaths by EU/EEA country                                                                        | Monitoring daily COVID-19 cases.                                                             | <a href="https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-COVID-19-eueea-country">https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-COVID-19-eueea-country</a> . [296]                                                                     | [296]      |
|                                                    | Data on SARS-CoV-2 variants in the EU/EEA                                                                                                                   | Monitoring SARS-CoV-2 variants in the EU/EEA                                                 | <a href="https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-COVID-19-eueea">https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-COVID-19-eueea</a> . [297]                                                                                       | [297]      |
|                                                    | Data on 14-day notification rate of new COVID-19 cases and deaths                                                                                           | Monitoring and analysis of Data on 14-day notification rate of new COVID-19 cases and deaths | <a href="https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-COVID-19">https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-COVID-19</a> . [298]                                                             | [298][299] |
|                                                    | Data on the daily subnational 14-day notification rate of new COVID-19 cases                                                                                | Monitoring and analysis.                                                                     | <a href="https://www.ecdc.europa.eu/en/publications-data/subnational-14-day-notification-rate-COVID-19">https://www.ecdc.europa.eu/en/publications-data/subnational-14-day-notification-rate-COVID-19</a> . [300]                                                                 | [300][301] |
|                                                    | Data on hospital and ICU admission rates and current occupancy for COVID-19                                                                                 | Monitoring and analysis.                                                                     | <a href="https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-COVID-19">https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-COVID-19</a> . [302] | [302][303] |
|                                                    | Data on country response measures                                                                                                                           | Monitoring and analysis.                                                                     | <a href="https://www.ecdc.europa.eu/en/publications-data/download-data-response-measures-COVID-19">https://www.ecdc.europa.eu/en/publications-data/download-data-response-measures-COVID-19</a> [304]                                                                             | [304]      |
|                                                    | Data on age-specific notification rate                                                                                                                      | Monitoring and analysis.                                                                     | <a href="https://www.ecdc.europa.eu/en/publications-data/COVID-19-data-14-day-age-notification-rate-new-cases">https://www.ecdc.europa.eu/en/publications-data/COVID-19-data-14-day-age-notification-rate-new-cases</a> . [305]                                                   | [305]      |
|                                                    | Data on council recommendations for mapping the coordinated approach to the restriction of free movement in response to the COVID-19 pandemic in the EU/EEA | Monitoring and analysis.                                                                     | <a href="https://www.ecdc.europa.eu/en/publications-data/indicators-maps-support-council-recommendation">https://www.ecdc.europa.eu/en/publications-data/indicators-maps-support-council-recommendation</a> . [306]                                                               | [306]      |
|                                                    | Historical data on the COVID-19 daily number of cases and deaths by country, worldwide                                                                      | Monitoring and analysis.                                                                     | <a href="https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-COVID-19-cases-worldwide">https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-COVID-19-cases-worldwide</a> . [307]                 | [307]      |
| Kaggle.com                                         | Daily information on the number of COVID-19                                                                                                                 | Monitoring and analysis.                                                                     | <a href="https://www.kaggle.com/sudalairajkumar/novel-coronavirus-2019-dataset">https://www.kaggle.com/sudalairajkumar/novel-coronavirus-2019-dataset</a> [308]                                                                                                                   | [308]      |

|                                                     |                                                                                                                                                                            |                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                         |       |
|-----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|
|                                                     | affected areas across the globe                                                                                                                                            |                                                                                                                 |                                                                                                                                                                                                                                                                                                                                                                         |       |
| World Health Organization                           | Information on country reported public measures to curb COVID-19.                                                                                                          | Monitoring and analysis                                                                                         | <a href="https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm">https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm</a> . [309]                                                                                                                                                                                                         | [309] |
| Johns Hopkins' electronic medical record, Epic      | Information on the patients that have been confirmed or are suspected of having contracted COVID-19                                                                        | For retrospective analysis of COVID-19 patient populations                                                      | <a href="https://ictr.johnshopkins.edu/coronavirus/jh-crown/">https://ictr.johnshopkins.edu/coronavirus/jh-crown/</a> [310]                                                                                                                                                                                                                                             | [310] |
| National Patient-Centered Clinical Research Network | Data model tracking insights on patients infected with COVID-19                                                                                                            | For understanding and defining demographics infected with SARS-CoV-2                                            | <a href="https://pcornet.org/news/pcornet-COVID-19-common-data-model-launched-enabling-rapid-capture-of-insights/">https://pcornet.org/news/pcornet-COVID-19-common-data-model-launched-enabling-rapid-capture-of-insights/</a> [311]                                                                                                                                   | [311] |
| Johns Hopkins COVID-19 collaboration platform       | Publicizing protocols whose PIs are open to various levels of collaboration.                                                                                               | Protocol collaboration.                                                                                         | <a href="https://covidcp.org/">https://covidcp.org/</a> . [312]                                                                                                                                                                                                                                                                                                         | [312] |
| National COVID Cohort Collaborative                 | Building a centralized national data resource that the research community can use to study COVID-19 and identify potential treatments as the pandemic continues to evolve. | Rapid collection and analysis of clinical, laboratory, and diagnostic data from hospitals and health care plans | <a href="https://ncats.nih.gov/n3c/about">https://ncats.nih.gov/n3c/about</a> . [313]                                                                                                                                                                                                                                                                                   | [313] |
| 4CE                                                 | COVID-19 positive cases and new death rates by country, overtime                                                                                                           | For data analysis and interpretation                                                                            | <a href="https://covidclinical.net/plots/paper-01/release-2020-04-11/dailycounts.html">https://covidclinical.net/plots/paper-01/release-2020-04-11/dailycounts.html</a> . [314]                                                                                                                                                                                         | [314] |
| 4CE                                                 | COVID-19 number of patients by country, by gender                                                                                                                          | For data analysis and interpretation                                                                            | <a href="https://covidclinical.net/plots/paper-01/release-2020-04-11/demographics.html">https://covidclinical.net/plots/paper-01/release-2020-04-11/demographics.html</a> [315]                                                                                                                                                                                         | [315] |
| 4CE                                                 | COVID-19 lab values corresponding to 14 LOINC Codes                                                                                                                        | For data analysis and interpretation                                                                            | <a href="https://covidclinical.net/plots/paper-01/release-2020-04-11/labs.html">https://covidclinical.net/plots/paper-01/release-2020-04-11/labs.html</a> [316]                                                                                                                                                                                                         | [316] |
| 4CE                                                 | Comparison of data from CSSE JHU                                                                                                                                           | For data analysis and interpretation                                                                            | <a href="https://covidclinical.net/plots/paper-01/release-2020-04-11/change.html">https://covidclinical.net/plots/paper-01/release-2020-04-11/change.html</a> . [317]                                                                                                                                                                                                   | [317] |
| 4CE                                                 | Participating sites visualized on maps                                                                                                                                     | For data analysis and interpretation                                                                            | <a href="https://covidclinical.net/plots/paper-01/release-2020-04-11/sites.html">https://covidclinical.net/plots/paper-01/release-2020-04-11/sites.html</a> [318]                                                                                                                                                                                                       | [318] |
| 4CE                                                 | Daily Count Data for International Electronic Health Record-Derived COVID-19 Clinical Course Profile                                                                       | For data analysis and interpretation                                                                            | <a href="https://figshare.com/articles/dataset/Daily_Count_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152976/1">https://figshare.com/articles/dataset/Daily_Count_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152976/1</a> . [319] | [319] |
| 4CE                                                 | Demographic data for International Electronic Health Record-Derived COVID-19 Clinical Course Profile.                                                                      | For data analysis and interpretation                                                                            | <a href="https://figshare.com/articles/dataset/Demographics_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152973/1">https://figshare.com/articles/dataset/Demographics_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152973/1</a> [320] | [320] |
| 4CE                                                 | Diagnosis data for International Electronic Health Record-Derived COVID-19 Clinical Course Profile.                                                                        | For data analysis and interpretation                                                                            | <a href="https://figshare.com/articles/dataset/Diagnosis_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152967">https://figshare.com/articles/dataset/Diagnosis_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152967</a> [321]           | [321] |
| 4CE                                                 | Labs data for International Electronic Health Record-Derived COVID-19 Clinical Course Profile.                                                                             | For data analysis and interpretation                                                                            | <a href="https://figshare.com/articles/dataset/Labs_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152766">https://figshare.com/articles/dataset/Labs_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152766</a> [322]                     | [322] |
| 4CE                                                 | Labs data for International Electronic Health Record-Derived COVID-19                                                                                                      | For data analysis and interpretation                                                                            | <a href="https://figshare.com/articles/dataset/Healthcare_Systems/12118911">https://figshare.com/articles/dataset/Healthcare_Systems/12118911</a> [323]                                                                                                                                                                                                                 | [323] |

|                                                                |                                                                                                |                                      |                                                                                                                                                                                                                                                                                                                                                                                 |            |
|----------------------------------------------------------------|------------------------------------------------------------------------------------------------|--------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
|                                                                | Clinical Course Profile.                                                                       |                                      |                                                                                                                                                                                                                                                                                                                                                                                 |            |
| 4CE                                                            | Time series COVID-19 confirmed cases                                                           | For data analysis and interpretation | <a href="https://github.com/CSSEGISandData/COVID-19/blob/dcd4181613f512a6f75249fc77b63286aeb7271/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv">https://github.com/CSSEGISandData/COVID-19/blob/dcd4181613f512a6f75249fc77b63286aeb7271/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv</a> [324] | [324]      |
| Health and Retirement Study                                    | 2020 HRS COVID-19 project                                                                      | For data analysis and interpretation | <a href="https://hrsdata.isr.umich.edu/data-products/2020-hrs-COVID-19-project">https://hrsdata.isr.umich.edu/data-products/2020-hrs-COVID-19-project</a> . [325]                                                                                                                                                                                                               | [325]      |
| COVID-19 research database                                     | Electronic health records, claims, and consumer data.                                          | For data analysis and interpretation | <a href="https://covid19researchdatabase.org/">https://covid19researchdatabase.org/</a> . [326]                                                                                                                                                                                                                                                                                 | [326]      |
| COVID-19 Research Initiatives in the HRS International Network | HRS COVID-19 Data on questionnaires, surveys, interviews, and state policies                   | For data analysis and interpretation | <a href="https://hrs.isr.umich.edu/data-products/COVID-19">https://hrs.isr.umich.edu/data-products/COVID-19</a> [327]                                                                                                                                                                                                                                                           | [327]      |
| Center for Disease Control and Prevention                      | COVID-19 Case Surveillance Public Use Data with Geography                                      | For analysis and interpretation      | <a href="https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4">https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4</a> . [328]                                                                                                                                                   | [328]      |
| Center for Disease Control and Prevention                      | COVID-19 Case Surveillance Public Use Data                                                     | For analysis and interpretation      | <a href="https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf">https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf</a> . [329]                                                                                                                                                                   | [329]      |
| Center for Disease Control and Prevention                      | COVID-19 Case Surveillance Restricted Access Detailed Data                                     | For analysis and interpretation      | <a href="https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t">https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detai/mbd7-r32t</a> . [330]                                                                                                                                                   | [330]      |
| Center for Disease Control and Prevention                      | COVID-19 Vaccine Distribution Allocations by Jurisdiction – Janssen                            | For analysis and interpretation      | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/w9zu-fywh">https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/w9zu-fywh</a> [331]                                                                                                                                                               | [331]      |
| Center for Disease Control and Prevention                      | COVID-19 Vaccine Distribution Allocations by Jurisdiction - Pfizer                             | For analysis and interpretation      | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg">https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg</a> [332]                                                                                                                                                               | [332][333] |
| Center for Disease Control and Prevention                      | United States COVID-19 Cases and Deaths by State over Time                                     | For analysis and interpretation      | <a href="https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36">https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36</a> [334]                                                                                                                                                     | [334]      |
| Center for Disease Control and Prevention                      | COVID-19 Vaccine Distribution Allocations by Jurisdiction – Moderna                            | For analysis and interpretation      | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws">https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws</a> . [335]                                                                                                                                                             | [335][336] |
| Center for Disease Control and Prevention                      | Provider Relief Fund COVID-19 Nursing Home Quality Incentive Program                           | For analysis and interpretation      | <a href="https://data.cdc.gov/Administrative/Provider-Relief-Fund-COVID-19-Nursing-Home-Quality/bfqg-cb6d">https://data.cdc.gov/Administrative/Provider-Relief-Fund-COVID-19-Nursing-Home-Quality/bfqg-cb6d</a> [337]                                                                                                                                                           | [337]      |
| Center for Disease Control and Prevention                      | Indicators of Anxiety or Depression Based on Reported Frequency of Symptoms During Last 7 Days | For analysis and interpretation      | <a href="https://data.cdc.gov/NCHS/Indicators-of-Anxiety-or-Depression-Based-on-Repor/8pt5-q6wp">https://data.cdc.gov/NCHS/Indicators-of-Anxiety-or-Depression-Based-on-Repor/8pt5-q6wp</a> [338]                                                                                                                                                                               | [338]      |
| Center for Disease Control and Prevention                      | Mental Health Care in the Last 4 Weeks                                                         | For analysis and interpretation      | <a href="https://data.cdc.gov/NCHS/Mental-Health-Care-in-the-Last-4-Weeks/yni7-er2q">https://data.cdc.gov/NCHS/Mental-Health-Care-in-the-Last-4-Weeks/yni7-er2q</a> [339]                                                                                                                                                                                                       | [339][340] |
| Center for Disease Control and Prevention                      | Vaccine Hesitancy for COVID-19: County and local estimate                                      | For analysis and interpretation      | <a href="https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw">https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw</a> . [341]                                                                                                                                                             | [341]      |
| Center for Disease Control and Prevention                      | Loss of Work Due to Illness from COVID-19                                                      | For analysis and interpretation      | <a href="https://data.cdc.gov/NCHS/Loss-of-Work-Due-to-Illness-from-COVID-19/qgkx-mswu">https://data.cdc.gov/NCHS/Loss-of-Work-Due-to-Illness-from-COVID-19/qgkx-mswu</a> . [342]                                                                                                                                                                                               | [342]      |
| Center for Disease Control and Prevention                      | COVID-19 Vaccinations in the United States by Jurisdiction                                     | For analysis and interpretation      | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/uns-k-b7fc">https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdi/uns-k-b7fc</a> . [343]                                                                                                                                                           | [343]      |
| Center for Disease Control and Prevention                      | Provider Relief Fund & Accelerated and Advance Payments                                        | For analysis and interpretation      | <a href="https://data.cdc.gov/Administrative/Provider-Relief-Fund-Accelerated-and-Advance-Payme/v2pi-w3up">https://data.cdc.gov/Administrative/Provider-Relief-Fund-Accelerated-and-Advance-Payme/v2pi-w3up</a> [344]                                                                                                                                                           | [344]      |
| Center for Disease Control and Prevention                      | Indicators of Reduced Access to Care Due to the Coronavirus Pandemic During Last 4 Weeks       | For analysis and interpretation      | <a href="https://data.cdc.gov/NCHS/Indicators-of-Reduced-Access-to-Care-Due-to-the-Co/xb3p-q62w">https://data.cdc.gov/NCHS/Indicators-of-Reduced-Access-to-Care-Due-to-the-Co/xb3p-q62w</a> . [345]                                                                                                                                                                             | [345]      |

|                                           |                                                                                                              |                                 |                                                                                                                                                                                                                                         |            |
|-------------------------------------------|--------------------------------------------------------------------------------------------------------------|---------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Center for Disease Control and Prevention | Access and Use of Telemedicine During COVID-19                                                               | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Access-and-Use-of-Telemedicine-During-COVID-19/8xy9-ubqz">https://data.cdc.gov/NCHS/Access-and-Use-of-Telemedicine-During-COVID-19/8xy9-ubqz</a> . [346]                                             | [346][347] |
| Center for Disease Control and Prevention | COVID-19 Vaccination Trends in the United States, National and Jurisdictional data                           | For analysis and interpretation | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2">https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2</a> [348]                       | [348]      |
| Center for Disease Control and Prevention | Reduced Access to Care During COVID-19                                                                       | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Reduced-Access-to-Care-During-COVID-19/th9n-ghnr">https://data.cdc.gov/NCHS/Reduced-Access-to-Care-During-COVID-19/th9n-ghnr</a> . [349]                                                             | [349]      |
| Center for Disease Control and Prevention | Telemedicine Use in the Last 4 Weeks                                                                         | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Telemedicine-Use-in-the-Last-4-Weeks/h7xa-837u">https://data.cdc.gov/NCHS/Telemedicine-Use-in-the-Last-4-Weeks/h7xa-837u</a> [350]                                                                   | [350][351] |
| Center for Disease Control and Prevention | Provisional COVID-19 Death Counts in the United States by County                                             | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-in-the-United-St/kn79-hsxy">https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-in-the-United-St/kn79-hsxy</a> [352]                                       | [352]      |
| Center for Disease Control and Prevention | Provisional COVID-19 Deaths: Focus on Ages 0-18 Years                                                        | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-Focus-on-Ages-0-18-Yea/nr4s-juj3">https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-Focus-on-Ages-0-18-Yea/nr4s-juj3</a> [353]                                       | [353]      |
| Center for Disease Control and Prevention | COVID-19 Vaccination and Case Trends by Age Group, United States                                             | For analysis and interpretation | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-and-Case-Trends-by-Age-Group-/gxj9-t96f">https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-and-Case-Trends-by-Age-Group-/gxj9-t96f</a> . [354]                     | [354]      |
| Center for Disease Control and Prevention | Excess Deaths Associated with COVID-19                                                                       | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Excess-Deaths-Associated-with-COVID-19/xkkf-xrst">https://data.cdc.gov/NCHS/Excess-Deaths-Associated-with-COVID-19/xkkf-xrst</a> [355]                                                               | [355]      |
| Center for Disease Control and Prevention | Indicators of Health Insurance Coverage at the Time of Interview                                             | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Indicators-of-Health-Insurance-Coverage-at-the-Tim/jb9g-gnvr">https://data.cdc.gov/NCHS/Indicators-of-Health-Insurance-Coverage-at-the-Tim/jb9g-gnvr</a> . [356]                                     | [356][357] |
| Center for Disease Control and Prevention | Provisional COVID-19 Death Counts by Week Ending Date and State                                              | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab">https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab</a> [359]                                       | [358][359] |
| Center for Disease Control and Prevention | COVID-19 Vaccination Demographics in the United States, National data                                        | For analysis and interpretation | <a href="https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Demographics-in-the-United-St/km4m-vcsb">https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Demographics-in-the-United-St/km4m-vcsb</a> [360]                       | [360]      |
| Center for Disease Control and Prevention | Nationwide Survey on Commercial Laboratory Seroprevalence                                                    | For analysis and interpretation | <a href="https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv">https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv</a> [361] | [361]      |
| Center for Disease Control and Prevention | Survey on COVID-19 Hospital Data from the National Hospital Care                                             | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/COVID-19-Hospital-Data-from-the-National-Hospital-/q3t8-zr7t">https://data.cdc.gov/NCHS/COVID-19-Hospital-Data-from-the-National-Hospital-/q3t8-zr7t</a> [362]                                       | [362][363] |
| Center for Disease Control and Prevention | Provisional COVID-19 Death Counts by Age in Years, 2020-2021                                                 | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Age-in-Years-/3apk-4u4f">https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Age-in-Years-/3apk-4u4f</a> [364]                                       | [364]      |
| Center for Disease Control and Prevention | Long-term Care and COVID-19                                                                                  | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Long-term-Care-and-COVID-19/3j26-kg6d">https://data.cdc.gov/NCHS/Long-term-Care-and-COVID-19/3j26-kg6d</a> [365]                                                                                     | [365]      |
| Center for Disease Control and Prevention | Provisional COVID-19 Deaths by Place of Death and State                                                      | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Place-of-Death-and-/uggs-hy5q">https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Place-of-Death-and-/uggs-hy5q</a> . [366]                                     | [366]      |
| Center for Disease Control and Prevention | Provisional COVID-19 Deaths by Week and Urbanicity                                                           | For analysis and interpretation | <a href="https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Week-and-Urbanicity/hkhe-f7hg">https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Week-and-Urbanicity/hkhe-f7hg</a> . [367]                                     | [367][368] |
| Center for Disease Control and Prevention | U.S. State and Territorial Stay-At-Home Orders: March 15, 2020 – August 15, 2021 by County by Day            | For analysis and interpretation | <a href="https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Stay-At-Home-Orders-Marc/y2iy-8irm">https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Stay-At-Home-Orders-Marc/y2iy-8irm</a> . [369]       | [369]      |
| Center for Disease Control and Prevention | U.S. State and Territorial Public Mask Mandates from April 10, 2020 through August 15, 2021 by County by Day | For analysis and interpretation | <a href="https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i">https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i</a> [370]         | [370][371] |

|                                            |                                                                                                                                 |                                                     |                                                                                                                                                                                                                                                                                                                                                                                                                           |            |
|--------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| Center for Disease Control and Prevention  | U.S. State, Territorial, and County Stay-At-Home Orders: March 15-May 5 by County by Day                                        | For analysis and interpretation                     | <a href="https://data.cdc.gov/Policy-Surveillance/U-S-State-Territorial-and-County-Stay-At-Home-Order/qz3x-mf9n">https://data.cdc.gov/Policy-Surveillance/U-S-State-Territorial-and-County-Stay-At-Home-Order/qz3x-mf9n</a> . [372]                                                                                                                                                                                       | [372]      |
| NCHS                                       | Provisional Death Counts for Influenza, Pneumonia, and COVID-19                                                                 | For analysis and interpretation                     | <a href="https://data.cdc.gov/NCHS/Provisional-Death-Counts-for-Influenza-Pneumonia-a/ynw2-4viq">https://data.cdc.gov/NCHS/Provisional-Death-Counts-for-Influenza-Pneumonia-a/ynw2-4viq</a> . [373]                                                                                                                                                                                                                       | [373][374] |
| European COVID-19 data platform            | Three data hubs reporting SARS-CoV-2, COVID-19, and Federated European Genome-phenome                                           | For data exploration, analysis, and interpretation. | <a href="https://www.covid19dataportal.org/the-european-COVID-19-data-platform">https://www.covid19dataportal.org/the-european-COVID-19-data-platform</a> [375]                                                                                                                                                                                                                                                           | [375]      |
| Open Safely                                | Computational resources and open access data to address COVID-19                                                                | For data exploration, analysis, and interpretation. | <a href="https://datascience.nih.gov/COVID-19-open-access-resources">https://datascience.nih.gov/COVID-19-open-access-resources</a> [376]                                                                                                                                                                                                                                                                                 | [376]      |
| ImmPort Shared Data                        | Research data available to the public and mostly scientific community to improve research work around COVID-19                  | For data exploration, analysis, and interpretation. | <a href="https://www.immport.org/shared/search?filters=study_2_condition_or_disease.condition_preferred:COVID-19%20-%20DOID:0080600&amp;utm_source=COVID-19&amp;utm_medium=banner&amp;utm_campaign=COVID-19">https://www.immport.org/shared/search?filters=study_2_condition_or_disease.condition_preferred:COVID-19%20-%20DOID:0080600&amp;utm_source=COVID-19&amp;utm_medium=banner&amp;utm_campaign=COVID-19</a> [377] | [377]      |
| World Health Organization                  | Global COVID-19 situation for confirmed cases.                                                                                  | For data exploration, analysis, and interpretation. | <a href="https://covid19.who.int/">https://covid19.who.int/</a> . [378]                                                                                                                                                                                                                                                                                                                                                   | [378]      |
| World meter                                | Global COVID-19 cases including confirmed cases, deaths, active cases, and closed cases.                                        | For data exploration, analysis, and interpretation. | <a href="https://www.worldometers.info/coronavirus/">https://www.worldometers.info/coronavirus/</a> [379]                                                                                                                                                                                                                                                                                                                 | [379]      |
| The World Bank                             | COVID-19 household monitoring dashboard.                                                                                        | For data exploration, analysis, and interpretation. | <a href="https://www.worldbank.org/en/data/interactive/2020/11/11/COVID-19-high-frequency-monitoring-dashboard">https://www.worldbank.org/en/data/interactive/2020/11/11/COVID-19-high-frequency-monitoring-dashboard</a> [380]                                                                                                                                                                                           | [380]      |
| The World Bank Group                       | COVID-19 business pulse survey dashboard that contains data on the socio-economic impacts of COVID-19 in 76 selected countries. | For data exploration, analysis, and interpretation. | <a href="https://www.worldbank.org/en/data/interactive/2021/01/19/COVID-19-business-pulse-survey-dashboard">https://www.worldbank.org/en/data/interactive/2021/01/19/COVID-19-business-pulse-survey-dashboard</a> . [381]                                                                                                                                                                                                 | [381][382] |
| The World Bank Group                       | Guidance to World Bank Group vendors on COVID-19.                                                                               | For data exploration, analysis, and interpretation. | <a href="https://www.worldbank.org/en/about/corporate-procurement/announcements/guidance_on_COVID-19">https://www.worldbank.org/en/about/corporate-procurement/announcements/guidance_on_COVID-19</a> [383]                                                                                                                                                                                                               | [383]      |
| The World Bank Group                       | Harmonized COVID-19 household monitoring survey                                                                                 | For data exploration, analysis, and interpretation. | <a href="https://datacatalog.worldbank.org/search/dataset/0037769/Harmonized-COVID-19-Household-Monitoring-Surveys">https://datacatalog.worldbank.org/search/dataset/0037769/Harmonized-COVID-19-Household-Monitoring-Surveys</a> . [384]                                                                                                                                                                                 | [384]      |
| Centers for Disease Control and Prevention | Effectiveness of COVID-19 vaccines                                                                                              | For data exploration, analysis, and interpretation. | <a href="https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness.html">https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness.html</a> . [385]                                                                                                                                                                                                                                                         | [385]      |
| Centers for Disease Control and Prevention | COVID-19 integrated country view                                                                                                | For data exploration, analysis, and interpretation. | <a href="https://covid.cdc.gov/covid-data-tracker/#county-view">https://covid.cdc.gov/covid-data-tracker/#county-view</a> . [386]                                                                                                                                                                                                                                                                                         | [386]      |
| Centers for Disease Control and Prevention | Forecasting cases and deaths COVID-19 in the United States                                                                      | For data exploration, analysis, and interpretation. | <a href="https://covid.cdc.gov/covid-data-tracker/#forecasting_weeklydeaths">https://covid.cdc.gov/covid-data-tracker/#forecasting_weeklydeaths</a> . [387]                                                                                                                                                                                                                                                               | [387]      |
| Centers for Disease Control and Prevention | COVID-19 vaccinations in the U.S                                                                                                | For data exploration, analysis, and interpretation. | <a href="https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total">https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total</a> [388]                                                                                                                                                                                                                                 | [388]      |
| Centers for Disease Control and Prevention | Country-level vulnerability index in the United States                                                                          | For data exploration, analysis, and interpretation. | <a href="https://covid.cdc.gov/covid-data-tracker/#pandemic-vulnerability-index">https://covid.cdc.gov/covid-data-tracker/#pandemic-vulnerability-index</a> [389]                                                                                                                                                                                                                                                         | [389]      |
| Centers for Disease Control and Prevention | COVID-19 community profile report                                                                                               | For data exploration, analysis, and interpretation. | <a href="https://healthdata.gov/Health/COVID-19-Community-Profile-Report/gqxm-d9w9">https://healthdata.gov/Health/COVID-19-Community-Profile-Report/gqxm-d9w9</a> [390]                                                                                                                                                                                                                                                   | [390]      |

## II. CONCLUSION

The summary tables (Tables I and II) present the technological resources and datasets used in tackling covid-19. Most of the data collected with COVID-19 related to hospitalizations, vaccinations, government response measures, deaths, confirmed reported cases, as well as restrictions and policies are used in aiding the pandemic. The R resources have mainly been used to develop Shiny apps and dashboards. Java, Kotlin, and Perl resources have been used in developing Android and iOS applications for contact tracing, disease surveillance, fast diagnostics, and notifying users anonymously if they have had any contact with someone who has been infected with COVID-19 via low-power bluetooth technology [28][31][32][51][52][69][129]. Based on the benefits of utilizing these resources, continued research and application of technological resources are highly recommendable [70].

## REFERENCES

- [1] <https://www.coronatracker.com/>
- [2] <https://github.com/JohnCoene/coronavirus.git>
- [3] A. Soetewey, "Top 100 R resources on Novel COVID-19 Coronavirus," Towards Data Science, 12 March 2020. [Online]. Available: <https://towardsdatascience.com/top-5-r-resources-on-covid-19-coronavirus-1d4c8df6d85f>. [Accessed 3 October 2021].
- [4] [https://chschoenenberger.shinyapps.io/covid19\\_dashboard/](https://chschoenenberger.shinyapps.io/covid19_dashboard/)
- [5] [https://github.com/chschoenenberger/covid19\\_dashboard](https://github.com/chschoenenberger/covid19_dashboard)
- [6] <https://nicohahn.shinyapps.io/covid19/>
- [7] [https://github.com/nicoFhahn/covid\\_shiny](https://github.com/nicoFhahn/covid_shiny)
- [8] <https://alhill.shinyapps.io/COVID19seir/>
- [9] [https://github.com/alsnhll/SEIR\\_COVID19](https://github.com/alsnhll/SEIR_COVID19)
- [10] <https://shubhrampandey.shinyapps.io/coronaVirusViz/>
- [11] <https://github.com/shubhrampandey/coronaVirus-dataViz>
- [12] <https://covid19forecast.science.unimelb.edu.au/>
- [13] <https://github.com/benflips/nCovForecast>
- [14] <https://dash.datascienceplus.com/covid19/>
- [15] <https://github.com/CSSEGISandData/COVID-19>
- [16] D. Anisa, "Map Visualization of COVID-19 Across the World with R," Data Science, 13 March 2020. [Online]. Available: <https://datascienceplus.com/map-visualization-of-covid19-across-world/>. [Accessed 3 October 2021].
- [17] <https://thibautfabacher.shinyapps.io/covid-19/>
- [18] <https://github.com/DrFabach/Corona>
- [19] <https://andrecalerovaldez.shinyapps.io/CovidTimeSeriesTest/>
- [20] <https://github.com/Sumidu/covid19shiny>
- [21] [https://tinu.shinyapps.io/Flatten\\_the\\_Curve/](https://tinu.shinyapps.io/Flatten_the_Curve/)
- [22] [https://github.com/tinu-schneider/Flatten\\_the\\_Curve](https://github.com/tinu-schneider/Flatten_the_Curve)
- [23] <https://jgassen.shinyapps.io/tidyCovid19/>
- [24] <https://statsandr.com/blog/top-r-resources-on-covid-19-coronavirus/#tidycovid19>
- [25] <https://sebastianwolf.shinyapps.io/Corona-Shiny/>
- [26] <https://github.com/zappingseb/coronashiny>
- [27] <https://c2m-africa.shinyapps.io/togo-covid-shiny/>
- [28] <http://modesty.securized.net/covid19prediction/>
- [29] <https://www.nathanchaney.com/>
- [30] N. Chaney, "Animating U.S. COVID-19 hotspots over time," NathanChaney, 23 October 2020. [Online]. Available: <https://www.nathanchaney.com/2020/10/09/animating-u-s-covid-19-hotspots-over-time/>. [Accessed 3 October 2021].
- [31] <https://c2m-africa.shinyapps.io/togo-covid-shiny/>
- [32] <https://c2m-africa.shinyapps.io/togo-covid-shiny/>
- [33] <http://modesty.usc.es:3838/covid19prediction/>
- [34] <https://github.com/arnimpdm/Covid-19-prediction>
- [35] <https://jontheepi.shinyapps.io/hcwcoronavirus/>
- [36] <https://github.com/jontheepi/hcwcoronavirus>
- [37] <https://rpubs.com/JMBodart/Covid19-hosp-be>
- [38] <https://github.com/jmbo1190/Covid19>
- [39] <https://covidminder.idea.rpi.edu/>
- [40] <https://github.com/TheRensselaerIDEA/COVIDMINDER>
- [41] IDEA, "COVIDMINDER: Revealing the regional disparities in outcomes, determinants, and mediations of the COVID-19 pandemic," RENSSELAER POLYTECHNIC INSTITUTE, 2021. [Online]. Available: <https://idea.rpi.edu/research/projects/covidminder>. [Accessed 04 October 2021].
- [42] [https://dataenthusiast.ca/apps/covid\\_ca/](https://dataenthusiast.ca/apps/covid_ca/)
- [43] <https://milano-r.github.io/erum2020-covidr-contest/petr-baranovskiy-covid-ca-data-explorer.html>
- [44] <https://jamalrogersapp.shinyapps.io/tsforecast/>
- [45] <https://github.com/fsmosca/COVID-19-PH-dataset>
- [46] <https://pharmhax.shinyapps.io/covid-corrector-shiny/>
- [47] <https://github.com/pharmhax/covid19-corrector>
- [48] K. M. Jagodnik, F. Ray, F. M. Giorgi and A. Lachmann, "Correcting under-reported COVID-19 case numbers: estimating the true scale of the pandemic," medRxiv., 2020.
- [49] <https://harpomaxx.shinyapps.io/covid19/>
- [50] <https://github.com/harpomaxx/COVID19>
- [51] <https://trafforddatalab.shinyapps.io/covid-19/>
- [52] <https://github.com/traffordDataLab/covid-19>
- [53] T. D. Lab, "Covid-19 resources.," Trafford, 2021. [Online]. Available: <https://www.trafforddatalab.io/covid19.html>. [Accessed 3 October 2021].
- [54] [https://trafforddatalab.shinyapps.io/trafford\\_covid-19/](https://trafforddatalab.shinyapps.io/trafford_covid-19/)
- [55] [https://github.com/traffordDataLab/trafford\\_covid-19](https://github.com/traffordDataLab/trafford_covid-19)
- [56] <https://covid-2019.live/en/>
- [57] <https://github.com/swsoyee/2019-ncov-japan>
- [58] <http://moduloinfo.ca/wordpress/>
- [59] <https://plugins.trac.wordpress.org/browser/covid-19-statistics-displayer>
- [60] <http://coronamapper.com/>
- [61] <https://github.com/JayWelsh/coronamap>
- [62] <https://petolau.shinyapps.io/coronadash/>
- [63] <https://github.com/PetoLau/CoronaDash>
- [64] <https://guillaumepressiat.shinyapps.io/covidfrance/>
- [65] <https://gist.github.com/GuillaumePressiat/0e3658624e42f763e3e6a67df92bc6c5>
- [66] <https://nicovidtracker.org/>
- [67] <https://github.com/YouGov-Data/covid-19-tracker>
- [68] Ulster University, "Ulster University Covid-19 tracker compares NI and ROI data on," Ulster University, 2020. [Online]. Available: <https://www.ulster.ac.uk/news/2020/june/ulster-university-covid-19-tracker-compares-ni-and-roi-data-on-coronavirus-testing,-positive-cases-and-deaths>. [Accessed 3 October 2021].
- [69] <https://scienceversuscorona.shinyapps.io/covid-overview/>
- [70] <https://github.com/fdabl/Covid-Overview>
- [71] <https://worldhealthorg.shinyapps.io/covid/>
- [72] <https://github.com/WorldHealthOrganization/app>
- [73] <https://www.covidsim.org/v6.20210915/>
- [74] <https://github.com/mrc-ide/squire>
- [75] <https://ramikrispin.github.io/coronavirus/>
- [76] <https://www.dhis2.org/>
- [77] <https://www.dhis2.org/>

- [78] Mobile phone download: <https://play.google.com/store/apps/details?id=com.dhis2&hl=en>
- [79] <https://github.com/dhis2/dhis2-covid19-doc>
- [80] Centre for Disease Control, "Guide to Global Digital Tools for COVID-19 Response," Centre for Disease Control and Prevention, 23 October 2020. [Online]. Available: <https://www.cdc.gov/coronavirus/2019-ncov/global-covid-19/compare-digital-tools.html>. [Accessed 8 October 2021].
- [81] <https://sormas.org/>
- [82] <https://github.com/hzi-braunschweig/SORMAS-Project>
- [83] <https://www.who.int/godata>
- [84] <https://github.com/godata-who/godata>
- [85] <https://www.cdc.gov/epiinfo/support/downloads.html>
- [86] <https://github.com/Epi-Info/Epi-Info-Community-Edition>
- [87] <https://getodk.org/software/>
- [88] <https://github.com/getodk/collect>
- [89] <https://www.dimagi.com/covid-19/>
- [90] <https://github.com/dimagi/commcare-hq>
- [91] <https://www.kobotoolbox.org>
- [92] <https://www.kobotoolbox.org>
- [93] <https://github.com/kobotoolbox>
- [94] <https://github.com/SHERLOCKLS/Detection-of-COVID-19-from-medical-images>
- [95] S. Liang, H. Liu, Y. Gu, X. Guo, L. H. L. Li and L. Tao, "Fast automated detection of COVID-19 from medical images using convolutional neural networks," *Communications Biology*, vol. 4, no. 1, pp. 1-13, 2021.
- [96] <https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker>
- [97] <https://github.com/OxCGRT/covid-policy-tracker>
- [98] Blavatnik School of Government, "Covid-19 Government Response Tracker," Blavatnik School of Government, 2021. [Online]. Available: <https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker>. [Accessed 3 October 2021].
- [99] <https://github.com/pcm-dpc/COVID-19>
- [100] <https://github.com/worldbank/covid-mobile-data>
- [101] <https://radarcovid.gob.es/>
- [102] <https://radarcovid.gob.es/>
- [103] <https://github.com/RadarCOVID/radar-covid-android>
- [104] van Dijk et al., "COVID RADAR app: Description and validation of population," *PLoS one.*, 2021.
- [105] <https://covidsafe.cs.washington.edu/>
- [106] <https://github.com/CovidSafe>
- [107] <https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html>
- [108] <https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html>
- [109] <https://github.com/cds-snc/covid-alert-app>
- [110] S.-A. P., "COVID Alert app cost feds \$20M but results 'did not meet expectations'," *Global News*, 2021. [Online]. Available: <https://globalnews.ca/news/8003920/covid-alert-app-expensive-ineffective/>. [Accessed 3 October 2021].
- [111] <https://erouska.cz/>
- [112] <https://github.com/covid19cz/erouska-android>
- [113] Data Proti Covid , "A joint activity of Czech technology companies and IT enthusiasts focused on helping in the fight against the COVID-19 infection," 2021. [Online]. Available: <https://covid19cz.cz/>. [Accessed 3 October 2021].
- [114] <https://www.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6>
- [115] <http://www.arcgis.com/apps/opsdashboard/index.html#/85320e2ea5424dfaaa75ae62e5c06e61>
- [116] [https://github.com/sidbannet/COVID-19\\_analysis](https://github.com/sidbannet/COVID-19_analysis)
- [117] <https://www.coronawarn.app/en/>
- [118] <https://github.com/corona-warn-app/cwa-app-android>
- [119] Corona-Warn-App, " Help us improve the Corona-Warn-App," 2021. [Online]. Available: <https://www.coronawarn.app/en/>. [Accessed 3 October 2021].
- [120] <https://www.tracetogether.gov.sg/>
- [121] <https://github.com/OpenTrace-Community>
- [122] Singapore Government Developer Portal, "TraceTogether – Community-driven Contact tracing," 2021. [Online]. Available: <https://www.developer.tech.gov.sg/technologies/digital-solutions-to-address-covid-19/tracetogether>. [Accessed 3 October 2021].
- [123] <https://www.health.govt.nz/our-work/diseases-and-conditions/covid-19-novel-coronavirus/covid-19-resources-and-tools/nz-covid-tracer-app>
- [124] <https://apps.apple.com/nz/app/nz-covid-tracer/id1511667597>
- [125] <https://github.com/minhealthnz/nzcovidtracer-app>
- [126] Ministry of Health New Zealand, "Open-source release of NZ COVID Tracer," 2020. [Online]. Available: <https://www.health.govt.nz/our-work/diseases-and-conditions/covid-19-novel-coronavirus/covid-19-resources-and-tools/nz-covid-tracer-app/about-nz-covid-tracer-app/open-source-release-nz-covid-tracer>. [Accessed 4 October 2021].
- [127] <https://github.com/dsaidgovsg/vigilantgantry>
- [128] Singapore Government Developer Portal, "VigilantGantry - Access Control with Artificial Intelligence (AI) and Video Analytics," 2021. [Online]. Available: <https://www.developer.tech.gov.sg/technologies/digital-solutions-to-address-covid-19/vigilantgantry>. [Accessed 4 October 2021].
- [129] <https://git-scm.com/downloads>
- [130] <https://github.com/sdsna/lancet-covid-19-database>
- [131] The Lancet , "COVID-19 Resource Centre," *Lancet*, 2021. [Online]. Available: <https://www.thelancet.com/coronavirus>. [Accessed 3 October 2021].
- [132] <https://art-bd.shinyapps.io/covid19canada/>
- [133] <https://github.com/ccodwg/Covid19Canada>
- [134] <https://github.com/mila-ijqa/COVI-ML>
- [135] <https://github.com/ImperialCollegeLondon/covid19model>
- [136] S. Flaxman, S. Mishra, A. Gandy, H. J. T. Unwin, T. A. Mellan, H. Coupland and S. ... Bhatt, "Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe," *Nature*, vol. 584, no. 7820, pp. 257-261., 2020.
- [137] <https://covidtracking.com/>
- [138] <https://github.com/COVID19Tracking>
- [139] The Covid-19 Tracking project, "How We Made the COVID Tracking Project," 2021. [Online]. Available: <https://covidtracking.com/>. [Accessed 3 October 2021].
- [140] <https://github.com/FedericoGarza/covidmx>
- [141] Federico, R., "covidmx: Python API to get information about COVID-19 in México.Python package version 0.3.1.," 2020. [Online]. Available: <https://github.com/FedericoGarza/covidmx>. [Accessed 3 October 2021].
- [142] <https://covid19-scenarios.org/>
- [143] [https://github.com/neherlab/covid19\\_scenarios](https://github.com/neherlab/covid19_scenarios)
- [144] <https://github.com/nhsx/covid-chest-imaging-database>
- [145] NHSx, "National COVID-19 Chest Imaging Database (NCCID).," National Health Service, 2021. [Online]. Available: <https://www.nhsx.nhs.uk/covid-19-response/data-and-covid-19/national-covid-19-chest-imaging-database-nccid/>. [Accessed 3 October 2021].
- [146] <https://www.nhsx.nhs.uk/covid-19-response/nhs-covid-pass-verifier-app/international-covid-pass-verifier-app-user-guide/>
- [147] <https://www.nhsx.nhs.uk/covid-19-response/nhs-covid-pass-verifier-app/international-covid-pass-verifier-app-user-guide/>
- [148] <https://github.com/nhsx/covid-pass-verifier>
- [149] <https://github.com/InstituteForDiseaseModeling/covasim>
- [150] K. e. al., "Determining the optimal strategy for reopening schools, the impact of test and trace interventions, and the risk of occurrence of a

- second COVID-19 epidemic wave in,” *The Lancet Child & Adolescent Health*, no. [https://doi.org/10.1016/S2352-4642\(20\)30250-9](https://doi.org/10.1016/S2352-4642(20)30250-9), 2020.
- [151] <https://covid-19-raffg.herokuapp.com/>
- [152] <https://github.com/raffg/covid-19>
- [153] <https://github.com/QuKunLab/COVID-19>
- [154] <https://github.com/ncbi-nlp/COVID-19-CT-CXR>
- [155] Y. Peng, Y. Tang, S. Lee, Y. Zhu, R. M. Summers and Z. Lu, “COVID-19-CT-CXR: a freely accessible and weakly labeled chest X-ray and CT image collection on COVID-19 from biomedical literature,” *IEEE transactions on big data*, pp. 2-12, 2020.
- [156] <https://github.com/covidcaremap/covid19-healthsystemcapacity>
- [157] <https://www.closedloop.ai/covid-19-index>
- [158] <https://github.com/closedloop-ai/cv19index>
- [159] Closed Loop Team, “Open-Source Data Science to Fight Covid-19,” 2021. [Online]. Available: <https://www.closedloop.ai/covid-19-index>. [Accessed 4 October 2021].
- [160] <https://github.com/BDI-pathogens/OpenABM-Covid19>
- [161] <https://pythonrepo.com/repo/pallupz-covid-vaccine-booking>
- [162] <https://pythonrepo.com/repo/pallupz-covid-vaccine-booking>
- [163] Mathieu, E., Ritchie, H., Ortiz-Ospina, E. et al. A global database of COVID-19 vaccinations. *Nat Hum Behav* (2021)
- [164] <https://ourworldindata.org/covid-vaccinations>.
- [165] <https://ourworldindata.org/covid-deaths>.
- [166] <https://ourworldindata.org/covid-cases>.
- [167] <https://ourworldindata.org/coronavirus-testing>.
- [168] <https://ourworldindata.org/covid-hospitalizations>.
- [169] <https://ourworldindata.org/mortality-risk-covid>.
- [170] <https://ourworldindata.org/mortality-risk-covid#citation>.
- [171] <https://ourworldindata.org/excess-mortality-covid>.
- [172] <https://ourworldindata.org/policy-responses-covid>.
- [173] <https://ourworldindata.org/identify-covid-exemplars>.
- [174] <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>.
- [175] <https://data.humdata.org/dataset/indonesia-covid-19-cases-recoveries-and-deaths-per-province>.
- [176] <https://data.humdata.org/dataset/coronavirus-covid-19-cases-and-deaths>.
- [177] <https://data.humdata.org/dataset/oxford-covid-19-government-response-tracker>.
- [178] <https://data.humdata.org/dataset/covid-19-vaccinations>.
- [179] <https://data.humdata.org/dataset/covid-19-global-travel-restrictions-and-airline-information>.
- [180] <https://data.humdata.org/dataset/nyt-covid-19-data>.
- [181] <https://data.humdata.org/dataset/total-covid-19-tests-performed-by-country>.
- [182] <https://data.humdata.org/dataset/global-school-closures-covid19>.
- [183] <https://data.humdata.org/dataset/fair-covid-dataset>.
- [184] <https://data.humdata.org/dataset/district-wise-quarantine-for-covid-19>.
- [185] <https://data.humdata.org/dataset/covid-19-data-visual-inputs>.
- [186] <https://data.humdata.org/dataset/corona-virus-covid-19-cases-and-deaths-in-venezuela>.
- [187] <https://data.humdata.org/dataset/ocha-global-humanitarian-operational-presence-who-what-where-3w-portal>.
- [188] <https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset>.
- [189] <https://data.humdata.org>
- [190] <https://data.humdata.org/dataset/europe-covid-19-subnational-cases>.
- [191] <https://data.humdata.org/dataset/philippines-covid-19-response-who-does-what-where>.
- [192] [https://data.humdata.org/dataset/open\\_one\\_time\\_covid\\_education\\_impact](https://data.humdata.org/dataset/open_one_time_covid_education_impact).
- [193] <https://data.humdata.org/dataset/google-mobility-report>.
- [194] [https://data.humdata.org/dataset/nigeria\\_covid19\\_subnational](https://data.humdata.org/dataset/nigeria_covid19_subnational).
- [195] <https://data.humdata.org/dataset/ecdc-covid-19>.
- [196] <https://data.humdata.org/dataset/immunization-campaigns-impacted>.
- [197] <https://data.humdata.org/dataset/financial-times-excess-mortality-during-covid-19-pandemic-data>.
- [198] <https://data.humdata.org/dataset/state-of-palestine-coronavirus-covid-19-subnational-cases>.
- [199] <https://data.humdata.org/dataset/covid-19-symptom-map>.
- [200] <https://data.humdata.org/dataset/covid-19-vaccine-doses-in-hrp-countries>.
- [201] <https://data.humdata.org/dataset/world-bank-indicators-of-interest-to-the-covid-19-outbreak>.
- [202] <http://globalhealth5050.org/covid19>.
- [203] <https://data.humdata.org/dataset/harmonized-covid-19-household-monitoring-surveys>.
- [204] [https://data.humdata.org/dataset/covid19\\_africa\\_continental\\_infections-recoveries-deaths](https://data.humdata.org/dataset/covid19_africa_continental_infections-recoveries-deaths).
- [205] <https://data.humdata.org/dataset/government-actions-on-covid-19>.
- [206] <https://data.humdata.org/dataset/preventive-health-survey>.
- [207] <https://data.humdata.org/dataset/cameroon-baseline-assessment-data-iom-dtm>.
- [208] <https://data.humdata.org/dataset/covax-round-6-allocations>.
- [209] [https://data.humdata.org/dataset/burkinafaso\\_covid19\\_subnational](https://data.humdata.org/dataset/burkinafaso_covid19_subnational).
- [210] <https://data.humdata.org/dataset/2019-novel-coronavirus-cases>.
- [211] <https://data.humdata.org/dataset/afghanistan-covid-19-statistics-per-province>.
- [212] <https://data.humdata.org/dataset/covid-19-mobility-italy>.
- [213] <https://data.humdata.org/dataset/africa-coronavirus-covid-19-subnational-cases>.
- [214] <https://data.humdata.org/dataset/covid-19-twitter-data-geographic-distribution>.
- [215] <https://data.humdata.org/dataset/global-covid-19-secondary-impacts>.
- [216] [https://data.humdata.org/dataset/burkinafaso\\_covid19\\_city-level](https://data.humdata.org/dataset/burkinafaso_covid19_city-level).
- [217] <https://data.humdata.org/dataset/peru-covid19-mortality-rate-in-lima>.
- [218] <https://data.humdata.org/dataset/covid-19-cases-data-in-moscow>.
- [219] <https://data.humdata.org/dataset/world-global-database-of-public-health-and-social-measures-applied-during-the-covid-19-pandemic>
- [220] <https://data.humdata.org/dataset/covid-19-impacts-africa>.
- [221] <https://data.humdata.org/dataset/iraq-coronavirus-covid-19-subnational-cases>.
- [222] <https://data.humdata.org/dataset/iati-covid19-funding>.
- [223] <https://data.humdata.org/dataset/world-bank-and-gavi-vaccine-financing>.
- [224] [https://data.humdata.org/dataset/open\\_one\\_time\\_covid\\_impact](https://data.humdata.org/dataset/open_one_time_covid_impact).
- [225] <https://data.humdata.org/dataset/ethiopia-covid19-cases>.
- [226] <https://data.humdata.org/dataset/covid-19-high-frequency-indicators>. [100]
- [227] <https://data.humdata.org/dataset/global-humanitarian-response-plan-covid-19-administrative-boundaries-and-population-statistics>.
- [228] <https://data.humdata.org/dataset/inform-covid-19-risk-index-version-0-1-4>.
- [229] <https://data.humdata.org/dataset/kazakhstan-coronavirus-covid-19-subnational-cases>.
- [230] <https://data.humdata.org/dataset/philippines-covid-19-operational-presence-risk-communication-and-community-engagement-rcece>.
- [231] <https://data.humdata.org/dataset/brazil-epidemiological-and-hospital-indicators-on-covid-19-in-ouro-preto>.
- [232] <https://data.humdata.org/dataset/oman-coronavirus-covid-19-subnational-cases>.
- [233] <https://data.humdata.org/dataset/mauritania-coronavirus-covid-19-city-level>.
- [234] <https://data.humdata.org/dataset/rapid-situation-tracking-for-covid-19-socioeconomic-impacts>.

- [235] <https://data.humdata.org/dataset/africa-covid-19-recovered-cases>.
- [236] <https://data.humdata.org/dataset/covid19-impacts-and-vaccine-acceptance-in-sub-saharan-africa>.
- [237] <https://data.humdata.org/dataset/un-collective-results-on-the-covid-19-socioeconomic-response-in-2020>.
- [238] <https://data.humdata.org/dataset/economic-impact-of-covid-19-in-sub-saharan-africa>.
- [239] <https://data.humdata.org/dataset/myanmar-coronavirus-covid-19-subnational-cases>.
- [240] <https://data.humdata.org/dataset/ghana-coronavirus-covid-19-subnational-cases>
- [241] <https://data.humdata.org/dataset/aid-security-and-covid-19>.
- [242] <https://data.humdata.org/dataset/ngos-affected-by-covid19-russia>.
- [243] <https://data.humdata.org/>
- [244] <https://data.humdata.org/dataset/covax-facility-interim-distribution-forecast>.
- [245] <https://data.humdata.org/dataset/unhcr-ken-2020-socioeconomic-impact-of-covid-19-on-pocs-in-kenya-v2-2>.
- [246] <https://data.humdata.org/dataset/togo-coronavirus-covid-19-city-level>.
- [247] <https://data.humdata.org/dataset/unicef-indicators-of-interest-to-the-covid-19-outbreak>.
- [248] <https://data.humdata.org/dataset/mozambique-coronavirus-covid-19-subnational-cases>.
- [249] <https://data.humdata.org/dataset/haiti-covid-19-subnational-cases>
- [250] <https://data.humdata.org/dataset/constraints-faced-by-people-due-to-covid-19-outbreak>.
- [251] <https://data.humdata.org/dataset/philippines-2020-significant-events>.
- [252] <https://data.humdata.org/dataset/unhcr-ken-2020-covid-round5-v2-1>.
- [253] <https://data.humdata.org/dataset/>
- [254] <https://data.humdata.org/dataset/sudan-covid-19-preparedness-and-response-4w>.
- [255] <https://data.humdata.org/dataset/community-feedback-by-indonesian-red-cross-pmi>.
- [256] <https://data.humdata.org/dataset/sudan-projected-covid-19-sub-national-cases>.
- [257] <https://data.humdata.org/dataset/travel-restriction-monitoring-iata-covid-19-iom-dtm>.
- [258] <https://data.humdata.org/dataset/west-and-central-africa-coronavirus-covid-19-situation>.
- [259] <https://data.humdata.org/dataset/somalia-projected-covid-19-subnational-cases>.
- [260] <https://data.humdata.org/dataset/ethiopia-coronavirus-covid-19-subnational-cases>.
- [261] <https://data.humdata.org/dataset/covid-19-cases-data-in-russia>.
- [262] <https://data.humdata.org/dataset/ongoing-impacts-of-covid-19-in-sub-saharan-africa>.
- [263] <https://data.humdata.org/dataset/covid-19-global-appeals-and-plans>.
- [264] <https://data.humdata.org/dataset/coronavirus-in-sub-saharan-africa>
- [265] <https://data.humdata.org/dataset/cerf-covid-19-allocations>.
- [266] <https://data.humdata.org/dataset/inform-covid-analysis-v01>.
- [267] <https://data.humdata.org/dataset/covid19-humanitarian-access>.
- [268] <https://data.humdata.org/dataset/lshrm-covid-19-projections>.
- [269] <https://data.humdata.org/dataset/eastern-and-southern-africa-covid-19-unicef-situation-and-response>.
- [270] <https://data.humdata.org/dataset/mali-coronavirus-covid-19-subnational-cases>.
- [271] <https://data.humdata.org/dataset/covid-19-economic-exposure-index>.
- [272] <https://data.humdata.org/dataset/somalia-coronavirus-covid-19-subnational-cases>.
- [273] <https://data.humdata.org/>
- [274] <https://data.humdata.org/>
- [275] <https://data.humdata.org/dataset/afghanistan-projected-covid-19-sub-national-cases>.
- [276] <https://data.humdata.org/dataset/unhcr-bgd-2020-covid-mwtk-v2-1>.
- [277] [https://data.humdata.org/dataset/covid19\\_risk\\_index\\_zip](https://data.humdata.org/dataset/covid19_risk_index_zip).
- [278] <https://data.humdata.org/dataset/democratic-republic-of-the-congo-coronavirus-covid-19-subnational-cases>.
- [279] <https://data.humdata.org/dataset/libya-coronavirus-covid-19-subnational-cases>.
- [280] <https://data.humdata.org/dataset/unhcr-ken-2020-socioeconomic-impact-of-covid-19-on-pocs-in-kenya-round2-v1-0>.
- [281] <https://data.humdata.org/dataset/cameroon-covid-19-mobility-restriction-point-of-entries-iom-dtm>.
- [282] <https://data.humdata.org/dataset/unhcr-ken-2020-covid-panel-v2-1>.
- [283] <https://data.humdata.org/dataset/iraq-projected-covid-19-sub-national-cases>.
- [284] <https://data.humdata.org/dataset/unhcr-nga-2020-sea-covid19-v2-1>.
- [285] <https://data.humdata.org/>
- [286] <https://data.humdata.org/dataset/ddi-zam-unhcr-covid19-impact-assessment-on-refugee-livelihoods-zambia-july-2020>.
- [287] <https://data.humdata.org/dataset/symptomatology-ecu911-santa-cruz-monthly-2018-2021>.
- [288] <https://data.humdata.org/>
- [289] <https://data.humdata.org/dataset/afghanistan-covid-19-health-facilities-by-province>.
- [290] <https://data.humdata.org/dataset/acaps-covid-19-humanitarian-exemptions-dataset>
- [291] <https://data.humdata.org/dataset/south-sudan-covid-19-mobility-and-preparedness-updates-iom-dtm>.
- [292] <https://data.humdata.org/dataset/unhcr-ken-2020-covid-round1-v2-2>.
- [293] <https://data.humdata.org/dataset/unhcr-ken-2020-covid-round4-v2-1>.
- [294] <https://data.humdata.org/dataset/unhcr-ken-2020-covid-round3-v2-1>.
- [295] <https://www.ecdc.europa.eu/en/publications-data/data-covid-19-vaccination-eu-eea>.
- [296] <https://www.ecdc.europa.eu/en/publications-data/data-daily-new-cases-COVID-19-eueea-country>.
- [297] <https://www.ecdc.europa.eu/en/publications-data/data-virus-variants-COVID-19-eueea>.
- [298] <https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-COVID-19>.
- [299] <https://www.ecdc.europa.eu/>
- [300] <https://www.ecdc.europa.eu/en/publications-data/subnational-14-day-notification-rate-COVID-19>.
- [301] <https://www.ecdc.europa.eu/>
- [302] <https://www.ecdc.europa.eu/en/publications-data/download-data-hospital-and-icu-admission-rates-and-current-occupancy-COVID-19>.
- [303] <https://www.ecdc.europa.eu/>
- [304] <https://www.ecdc.europa.eu/en/publications-data/download-data-response-measures-COVID-19>
- [305] <https://www.ecdc.europa.eu/en/publications-data/COVID-19-data-14-day-age-notification-rate-new-cases>.
- [306] <https://www.ecdc.europa.eu/en/publications-data/indicators-maps-support-council-recommendation>.
- [307] <https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-COVID-19-cases-worldwide>.
- [308] <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset>
- [309] <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/phsm>.
- [310] <https://ictr.johnshopkins.edu/coronavirus/jh-crown/>
- [311] <https://pcornet.org/news/pcornet-COVID-19-common-data-model-launched-enabling-rapid-capture-of-insights/>
- [312] <https://covidcp.org/>.
- [313] <https://ncats.nih.gov/n3c/about>.

- [314] <https://covidclinical.net/plots/paper-01/release-2020-04-11/dailycounts.html>.
- [315] <https://covidclinical.net/plots/paper-01/release-2020-04-11/demographics.html>
- [316] <https://covidclinical.net/plots/paper-01/release-2020-04-11/labs.html>
- [317] <https://covidclinical.net/plots/paper-01/release-2020-04-11/change.html>.
- [318] <https://covidclinical.net/plots/paper-01/release-2020-04-11/sites.html>
- [319] [https://figshare.com/articles/dataset/Daily\\_Count\\_Data\\_for\\_International\\_Electronic\\_Health\\_Record-Derived\\_COVID-19\\_Clinical\\_Course\\_Profile\\_The\\_4CE\\_Consortium/12152976/1](https://figshare.com/articles/dataset/Daily_Count_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152976/1).
- [320] [https://figshare.com/articles/dataset/Demographics\\_Data\\_for\\_International\\_Electronic\\_Health\\_Record-Derived\\_COVID-19\\_Clinical\\_Course\\_Profile\\_The\\_4CE\\_Consortium/12152973/1](https://figshare.com/articles/dataset/Demographics_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152973/1)
- [321] [https://figshare.com/articles/dataset/Diagnosis\\_Data\\_for\\_International\\_Electronic\\_Health\\_Record-Derived\\_COVID-19\\_Clinical\\_Course\\_Profile\\_The\\_4CE\\_Consortium/12152967](https://figshare.com/articles/dataset/Diagnosis_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152967)
- [322] [https://figshare.com/articles/dataset/Labs\\_Data\\_for\\_International\\_Electronic\\_Health\\_Record-Derived\\_COVID-19\\_Clinical\\_Course\\_Profile\\_The\\_4CE\\_Consortium/12152766](https://figshare.com/articles/dataset/Labs_Data_for_International_Electronic_Health_Record-Derived_COVID-19_Clinical_Course_Profile_The_4CE_Consortium/12152766)
- [323] [https://figshare.com/articles/dataset/Healthcare\\_Systems/12118911](https://figshare.com/articles/dataset/Healthcare_Systems/12118911)
- [324] [https://github.com/CSSEGISandData/COVID19/blob/dcd4181613f512a6f75249fc77b63286a6be7271/csse\\_covid\\_19\\_data/csse\\_covid\\_19\\_time\\_series/time\\_series\\_covid19\\_confirmed\\_global.csv](https://github.com/CSSEGISandData/COVID19/blob/dcd4181613f512a6f75249fc77b63286a6be7271/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv)
- [325] <https://hrsdata.isr.umich.edu/data-products/2020-hrs-COVID-19-project>.
- [326] <https://covid19researchdatabase.org/>.
- [327] <https://hrs.isr.umich.edu/data-products/COVID-19>
- [328] <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data-with-Ge/n8mc-b4w4>.
- [329] <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Public-Use-Data/vbim-akqf>.
- [330] <https://data.cdc.gov/Case-Surveillance/COVID-19-Case-Surveillance-Restricted-Access-Detail/mbd7-r32t>.
- [331] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/w9zu-fywh>
- [332] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg>
- [333] [205] <https://data.cdc.gov/>
- [334] <https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36>
- [335] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws>.
- [336] <https://data.cdc.gov/>
- [337] <https://data.cdc.gov/Administrative/Provider-Relief-Fund-COVID-19-Nursing-Home-Quality/bfqg-cb6d>
- [338] <https://data.cdc.gov/NCHS/Indicators-of-Anxiety-or-Depression-Based-on-Report/8pt5-q6wp>
- [339] <https://data.cdc.gov/NCHS/Mental-Health-Care-in-the-Last-4-Weeks/yini7-er2q>
- [340] <https://data.cdc.gov/>
- [341] <https://data.cdc.gov/Vaccinations/Vaccine-Hesitancy-for-COVID-19-County-and-local-es/q9mh-h2tw>.
- [342] <https://data.cdc.gov/NCHS/Loss-of-Work-Due-to-Illness-from-COVID-19/qgkx-mswu>.
- [343] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-Jurisdiction/uns-kb7fc>.
- [344] <https://data.cdc.gov/Administrative/Provider-Relief-Fund-Accelerated-and-Advance-Payment/v2pi-w3up>
- [345] <https://data.cdc.gov/NCHS/Indicators-of-Reduced-Access-to-Care-Due-to-the-Co/xb3p-q62w>.
- [346] <https://data.cdc.gov/NCHS/Access-and-Use-of-Telemedicine-During-COVID-19/8xy9-ubqz>.
- [347] <https://data.cdc.gov/>
- [348] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Trends-in-the-United-States-N/rh2h-3yt2>
- [349] <https://data.cdc.gov/NCHS/Reduced-Access-to-Care-During-COVID-19/th9n-ghnr>.
- [350] <https://data.cdc.gov/NCHS/Telemedicine-Use-in-the-Last-4-Weeks/h7xa-837u>
- [351] <https://data.cdc.gov/>
- [352] <https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-in-the-United-States/kn79-hsxy>
- [353] <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-Focus-on-Ages-0-18-Year/nr4s-juj3>
- [354] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-and-Case-Trends-by-Age-Group/gxj9-t96f>.
- [355] <https://data.cdc.gov/NCHS/Excess-Deaths-Associated-with-COVID-19/xkkf-xrst>
- [356] <https://data.cdc.gov/NCHS/Indicators-of-Health-Insurance-Coverage-at-the-Time/jb9g-gnvr>.
- [357] <https://data.cdc.gov/>
- [358] <https://data.cdc.gov/>
- [359] <https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Week-Ending-D/r8kw-7aab>
- [360] <https://data.cdc.gov/Vaccinations/COVID-19-Vaccination-Demographics-in-the-United-States/km4m-vcsb>
- [361] <https://data.cdc.gov/Laboratory-Surveillance/Nationwide-Commercial-Laboratory-Seroprevalence-Su/d2tw-32xv>
- [362] <https://data.cdc.gov/NCHS/COVID-19-Hospital-Data-from-the-National-Hospital/q3t8-zr7t>
- [363] <https://data.cdc.gov/>
- [364] <https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Age-in-Years-/3apk-4u4f>
- [365] <https://data.cdc.gov/NCHS/Long-term-Care-and-COVID-19/3j26-kg6d>
- [366] <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Place-of-Death-and-/uggs-hy5q>.
- [367] <https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Week-and-Urbanicity/hkhe-f7hg>.
- [368] <https://data.cdc.gov/>
- [369] <https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Stay-At-Home-Orders-Marc/y2iy-8irm>.
- [370] <https://data.cdc.gov/Policy-Surveillance/U-S-State-and-Territorial-Public-Mask-Mandates-Fro/62d6-pm5i>.
- [371] <https://data.cdc.gov/>
- [372] <https://data.cdc.gov/Policy-Surveillance/U-S-State-Territorial-and-County-Stay-At-Home-Orders/qz3x-mf9n>.
- [373] <https://data.cdc.gov/NCHS/Provisional-Death-Counts-for-Influenza-Pneumonia-a/ynw2-4viq>.
- [374] <https://data.cdc.gov/NCHS/>
- [375] <https://www.covid19dataportal.org/the-european-COVID-19-data-platform>.
- [376] <https://datascience.nih.gov/COVID-19-open-access-resources>.
- [377] [https://www.immport.org/shared/search?filters=study\\_2\\_condition\\_or\\_disease.condition\\_preferred:COVID-19%20-%20DOI:0080600&utm\\_source=COVID-19&utm\\_medium=banner&utm\\_campaign=COVID-19](https://www.immport.org/shared/search?filters=study_2_condition_or_disease.condition_preferred:COVID-19%20-%20DOI:0080600&utm_source=COVID-19&utm_medium=banner&utm_campaign=COVID-19).
- [378] <https://covid19.who.int/>.
- [379] <https://www.worldometers.info/coronavirus/>.
- [380] <https://www.worldbank.org/en/data/interactive/2020/11/11/COVID-19-high-frequency-monitoring-dashboard>.
- [381] <https://www.worldbank.org/en/data/interactive/2021/01/19/COVID-19-business-pulse-survey-dashboard>.
- [382] <https://www.worldbank.org>.
- [383] [https://www.worldbank.org/en/about/corporate-procurement/announcements/guidance\\_on\\_COVID-19](https://www.worldbank.org/en/about/corporate-procurement/announcements/guidance_on_COVID-19).
- [384] <https://datacatalog.worldbank.org/search/dataset/0037769/Harmonized-COVID-19-Household-Monitoring-Surveys>.
- [385] <https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness.html>.

- [386]<https://covid.cdc.gov/covid-data-tracker/#county-view>.
- [387][https://covid.cdc.gov/covid-data-tracker/#forecasting\\_weeklydeaths](https://covid.cdc.gov/covid-data-tracker/#forecasting_weeklydeaths).
- [388][https://covid.cdc.gov/covid-data-tracker/#vaccinations\\_vacc-total-admin-rate-total](https://covid.cdc.gov/covid-data-tracker/#vaccinations_vacc-total-admin-rate-total).
- [389]<https://covid.cdc.gov/covid-data-tracker/#pandemic-vulnerability-index>.
- [390]<https://healthdata.gov/Health/COVID-19-Community-Profile-Report/gqxm-d9w9>.
- [391]Muniswamaiah, Manoj, Tilak Agerwala, and Charles C. Tappert. "Survey of the use of digital technologies to combat COVID-19." In 2020 IEEE International Conference on Big Data (Big Data), pp. 5768-5771. IEEE, 2020.
- [392]Hannah Ritchie, Edouard Mathieu, Lucas Rod s-Guirao, Cameron Appel, Charlie Giattino, Esteban Ortiz-Ospina, Joe Hasell, Bobbie Macdonald, Diana Beltekian and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: '<https://ourworldindata.org/coronavirus>' (Online Resource).
- [393]Hasell, J., Mathieu, E., Beltekian, D. et al. A cross-country database of COVID-19 testing. *Sci Data* 7, 345 (2020)

# A Comparison between Online and Offline Health Seeking Information using Social Networks for Patients with Chronic Health Conditions

Dr Andrew Kear<sup>1</sup>

Faculty of Media and Communication  
Bournemouth University, UK

Simon Talbot<sup>2</sup>

Welspect Healthcare  
UK

**Abstract**—The patient is now better connected with other patients just like the consumer is now better connected with other consumers in particular through the growing adoption of social media and online peer to peer communities. These relationships which become collaborative have either positive or indeed negative consequences that may either endorse or have implications for a firm's products [32]. The aim of this research was to gain an understanding of the impact social media has on patient influence on healthcare provision especially in relation to information seeking and clinical product choice. It compares a group of patients who are predominantly online information seekers with a group who are predominantly offline information seekers. Bias will be eliminated by utilising probability sampling techniques in order to be able to perform statistical analysis on the results obtained. This study capitalises on having access to approximately 8000+ Direct to Patient consumers who are currently receiving devices for the management of their bladder problems. The intention of this research project is to gain an understanding of how two way online interactions have developed between patients with similar chronic medical conditions and how firms can use online social media to improve their relationship with patients. The key research question of this paper is: Have online social media tools affected demand for healthcare intermediation in patients, who experience chronic medical conditions and reflect a need to become better informed. The findings of this pre-Covid research were that, for patient groups that had chronic conditions, there was a positive relationship between time spent in developed peer to peer communities, are more trusting of online information and spend more time online.

**Keywords**—Component; social media; healthcare; peer to peer networks; patient networks; pre-Covid

## I. INTRODUCTION

The digital revolution through information technology has had an empowering influence on interactions between consumers (the end users) and consumers, and consumers and the marketplace. Social media has allowed instant reach to and sharing of information [14] and Social media penetration world-wide is continuously on the uptrend to move to over 3 billion by 2021 (Statista.com) [34]. Social media interactions offer a medium to be used by patients with informational, emotional, and social support pertaining to their issues [42]. This can offer both valuable information and in some cases misleading information. Consumers, in this case patients, have an increasingly loud voice and identity as online social peer to

peer groups form, 9 members of those groups or communities interact with each other [29] and social networks have now become a major component of popular culture [4]. Users of member communities are potentially open to influences, especially from other members within these communities. To search for other patients (consumers) knowledge and experiences [1] remains a key driver in this context. Individuals who never meet face to face today develop an identity and presence within member communities spanning across the globe. Facebook allowed patients to follow health-related pages. YouTube and Twitter are the next two widely used social media platforms. Patients are also able to participate in disease-specific group discussions [2] In a McKinsey survey (2011) involving 4261 respondents 67% of Healthcare companies use at least one social technology tool [16]. Now most businesses, social causes, political movements, public figures and governments' attempt to harness the power of social network sites such as Facebook due to the level of exposure and influence it offers [4]. Thus the benefits of adopting social media for the firm are clear and it is proposed that the patient-provider relationship is enhanced across all age groups (Ybarra & Suman, 2008). One of the key challenges faced by firms, according to a 2012 McKinsey survey, is a lack of quality detailed customer data, for example consumer interests or attitudes [5]. This in turn means that marketing decisions have to be based on known or 'comfortable' data such as internal sales data. This McKinsey survey concluded that less than 20% of marketing decisions are based on external quantitative and qualitative data such as consumer insights, these insights were described as difficult to obtain due to the fact that they are not readily available to the firm [5].

Information seeking is linked to decision making. Roxane Divol [28], in her article entitled 'Demystifying Social Media' describes the consumer decision journey as consisting of 5 stages: 'Consider, evaluate, purchase, experience then advocate' [28]. This stage process has resemblances with the search, experience, and credence model that has been studied in on line shopping [11]. The author in [28] in her paper argues that it is much easier to interact with the consumer at each of the five decision making stages through social media, compared with a more traditional paid media marketing strategy. This means that a social media marketing strategy brings undoubted benefits such as access to the consumer at each stage of the decision making journey. However social

marketing is of course not without its risks. Online communities such as blogs and peer to peer user groups can campaign for unproven or untested treatments. One example is a new untested surgical treatment for Multiple Sclerosis being 'advocated' by an online Canadian community group [10]. Today most clinicians take a positive view of the information found through online searches that patients bring to their clinics [36].

According to [38] 'patients are acquiring more power in the health care supply chain and their preferences are influencing manufacturers, physicians and hospitals' [36]. Therefore patients have influence over prescribed products they receive and have power over which products will appear on formularies in future. Online health information has now grown to become one of the most important information sources for people [17]. Pharmaceutical companies meanwhile provide product information that increases request rates [24]. Patients with chronic health conditions carry out regular searching for information relating to new treatments, nutritional advice and alternative therapies. Patients use both online communities and chat rooms. Today the adoption of online health information seeking behaviour is creating a more informed patient. This greater transparency of information does create issues. For instance, the large number of products potentially creates confusion amongst clinicians which in turn leads to examples of products being unnecessarily or inappropriately prescribed. A key tactic amongst newly formed Clinical Commissioning Groups is to develop what is known as a 'formulary', essentially this is a list limiting the prescription of products. These formularies limit the availability of some of these devices, based on criteria such as clinical or cost effectiveness.

Key questions arise as to differences between online and offline health seeking behaviours and their implications for health care provision. If a proven difference exists between online and off-line information seeking behaviours, especially in the context of patients with long term chronic illnesses, then this would have significant implications for the understanding of consumer behaviour. It would either imply that the patient had already 'decided' before they consulted with a clinician or at least would suggest that there are external influences on choice of treatment that the patient receives from their clinician. A further thing to note is that there are many benefits to researching illness and chronic illnesses beyond that of metaphor [33].

The factors influencing peer to peer communities can be grouped into the following categories:

**Healthcare Professionals** - Declining appointment times are typical; many patients are less satisfied with information and support obtained from clinicians [3]. Poor communication with physicians and the patient physician relationship may drive users to seek online channels for healthcare information [40]. Diffusing health information has a negative effect on frequency of health care/doctor visits [36].

**Behavioural Factors** - Online healthcare Information seeking behaviour is associated with healthier people [3]. Perceived poor health status positively affects both frequency and diversity of search for online health information [40].

Chronic disease sufferers whose condition worsens may look for health information online more often overtime and therefore have less doctor visits [36].

It is noteworthy that offline conversations also take place. A study that was conducted by [31] and discusses information seeking in women before visiting their GP and found that depending on whether these conversations were with 'kin' or 'friends', a difference in frequency of GP visits was noted. Essentially the conversations taking place within kinship networks were more intense and resulted in increased GP consultations [31]. The translation of this research to the social media era could be indicative of differing behaviour amongst diverse social groups that transcends the technology. However it does not always follow that patients who have increased healthcare information needs will seek more doctor visits as highlighted by [36].

[25] suggested that the internet is a key influence in changing the balance of 'power' between healthcare professionals and the public. The patient is becoming more knowledgeable and involved in health care decision-making and this is contributing to the de-professionalising of medicine [25]. This professional practice has often related to knowledgeability or expertise. [3] Highlights paternalistic attitudes of some doctors and nurses and the fact that the internet has created a digital divide providing an opportunity for educated, wealthier people to seek alternative healthcare opinions [3]. In particular the use of social networking has also grown from 5% of all adults in 2005, to 50% of all adults by 2011 [43]. According to a health research Institute survey in 2012 one-third of consumers are now using social media for 'health related matters'; preferring community sites over sponsored sites [17]. This implies that patients do not necessarily have to be in poor health to be motivated enough to seek healthcare information. According to the [20] Survey, 39% of consumers who have a broadband connection at home reported that they had used the internet in the last week for 'finding health information', this figure has increased by 3% since 2011. Finding health information was highlighted as having the most marked increase since last year indicating a growing trend towards the use of the internet for this purpose [20]. Similarly according to the 'Pew Internet and American Life Project' research, 35% of Americans say that at one time or another they have gone online specifically to try to find out what medical condition they or someone else might have. People are increasingly using 'ask a doctor' sites, for example 8% of internet users say they have in the last 12 months posted a health related question online [35]. This research adds further strength to the argument that the internet has increasingly become a compliment to formal healthcare information provision. However a consideration remains that is the law of e-healthcare attrition [6]. In addition some issues remain in relation to the authenticity verification by online health information seekers [8] in addition to the perceived credibility of the internet varied because expertise and trustworthiness were sometimes difficult to determine [12]. Additional consideration is that of the challenge to the authority of the expert causing a perceived deterioration in the physician-patient relationship [19].

This literature review captures two themes. First 'Increased healthcare information seeking behaviour will positively influence the patient's healthcare choices'. Several authors have contributed to the idea that patients do have influence over their healthcare choices [13], [36] including compliance with the advice. The question arises as to whether online information seeking has the same influence on patients with chronic health conditions and furthermore how known demographic factors such as gender, age and health status are influencing this.

The second theme captured by the literature review is 'Are patients seeking healthcare information online becoming more demanding of their healthcare compared with patients seeking healthcare information offline?' [10] discussed how social media messages can spread rapidly and influence demand. It is undetermined whether people seek more clinical appointments as a consequence [31], [36]. The other outcome of patients becoming more self efficacious is a possible reduction in reliance on expert intermediaries, and a consequent reduction in requests for healthcare appointments. A, dynamic Intermediation-Disintermediation-Apomediation (D.I.D.A) model has been proposed for this process [7]. The key question is whether online healthcare information searches are simply meeting a demand for personal understanding of the situation they are in, or if information seeking drives patients to seek more doctor appointments to obtain specific treatments.

The following objectives have therefore been developed for this research.

- 1) To find out which social networks most often used by patients with chronic health conditions.
- 2) To evaluate how patients with chronic health conditions utilise online healthcare information.
- 3) To find out how healthcare companies can better engage with patients who live with chronic health conditions.
- 4) To determine demographic differences in adoption of online social media tools used by patients with chronic health conditions.

## II. METHODOLOGY

The patient population studied in this research project consists of a sample from 8000+ patients who have a chronic health condition. Approximately 11% of these patients are cared for by another individual, either their parent or spouse. A survey was conducted in 2013 as part of a Direct to Patient prescription service audit; this was a requirement by the Department of Health. The resulting patient demographic profile was obtained based on this survey; there were 258 respondents.

The survey indicates that nearly 2/3 of the patients are over 65 years of age. One key factor that determines the degree of internet usage and adoption is age. This might suggest that these patients are relatively low social network users. One-third of the patients in this survey were female, and according to [36] being female increases your likelihood for online seeking of health information. However [41] found that being middle aged was associated with increased internet use in the

context of health care [41]. Taken together this research suggests that the population under study will have a good mix of online and off-line information seekers.

The analysis that follows uses correlation coefficients to test the relationship between two variables [18]. These variables are either positively or negatively correlated and correlations can range in their strength from weak to strong. For this research project weak correlations are taken as below 0.35 (positively or negatively) and correlations above 0.65 (positive or negative) are taken as strong (see, for example, [18]). Correlation analysis explores the extent to which two variables are related to each other without assuming causality. The purpose of performing a correlation is to enable a prediction about one variable based on what is known about another variable; if two variables are strongly correlated then a prediction of the movement of one can be based on the other [18].

For this research project the chosen method will need to produce results that are highly replicable if an understanding is to be gained about a wider chronic long term patient population and their behaviours. To be able to make management decisions relating to the findings of this research a statistical approach will be needed [30]. Hence the quantitative approach will be chosen. The requirement for a large and relatively easy to reach sample frame further supports the quantitative approach; as the 8000+ patient database allows for a relatively large patient sample to be taken. Therefore the quantitative research method represents the most opportunistic and practicable method. For this research project the probability sampling method that will be chosen is systematic sampling because this is simpler to implement. From the database of 8000+ patients an initial sampling frame will be selected consisting of patients with chronic conditions only. This excludes short term patients and allows for a systematic sampling method to select a more manageable sample before conducting a survey.

Focusing on a specific disease or condition can produce patients that are not representative of the population. This is because not everyone with a condition has consulted a doctor or specialist about their condition. This might be an important consideration in the context of this research project because patients who seek online or offline healthcare information before seeing the clinician may satisfy their information needs and therefore choose not to see the doctor or specialist. The sample here will only include chronic or long term patients that have consulted with a doctor or specialist thus introducing a degree of bias. Both time and cost constraints prohibit being able to sample patients who have not visited clinicians, as discussed already.

## III. ANALYSIS AND FINDINGS

The 8000+ database was screened for respondents who were greater than 16 years old and who had ordered products recently meaning that they were current users. Each of the respondents was checked to ensure that they were long-term users by showing that their date of first registration, on the database, was greater than 12 months. This process produced a sampling frame containing 3,767 records.

Using a stratified sampling technique, where every 6<sup>th</sup> patient was selected, a sample of approximately 628 subjects was produced and then targeted with the questionnaire. One hundred and forty five questionnaires were completed with an additional 13 responses which were deemed incomplete. The overall response rate was 25%.

The majority of the sample was from the 55+ age brackets with 25% aged between 55 and 64 and 52% aged 65+. Table I shows the percentage split of the research sample in comparison with the DOH sample. The gender split across each of the age bands was also fairly evenly split except for the 65+ age band which was male dominated, possibly reflecting disease specific demographic trends such as prostate related conditions in older males.

There was some general hypothesis tested using correlation analysis and t Test for independent samples. regarding age, self-efficacy and time within this study.

The following are the hypothesis, results and related discussion thereof.

Hypothesis H1A: Age has a positive effect on health information seeking behaviour online

This was tested against high seeking versus low seeking health information. The correlation analysis suggested that there was a negative correlation between age and health information seeking behaviour meaning that as age increased online information seeking decreased. The t-test performed in this analysis suggested that there was not a significant difference ( $p = 0.068$ ) between those patients classified as high information seekers and those classified as low information seekers online. There was not enough evidence to reject the null hypothesis. As such this was a non-significant finding.

According to [3] age, education and wealth were all associated with increased Internet use. Age was a key factor that discriminated between online and off-line information seekers [3]. The internet adoption rate for 65+ year old patients, the group most represented in this research sample, was twice the level found by the Office for National Statistics (20% vs 10%) [21]. This suggests that patients with chronic health conditions are more likely to use the internet compared to the general population. This did not seem to overcome the negative correlation with age. [31] described a measure of demand as an increase in clinical consultations. What therefore is a true measure of demand? The measure for demand must depend on what outcomes patients expect as a result of healthcare information seeking. This research project repeated [31] work with kinship (off-line) versus friendship networks being compared. However as the number of online general Internet users is so high in this research project (see Fig. 3) it is difficult to screen for 'off-line only' users. The 'Non Significant Finding' therefore may not be a surprise. One of the limitations of this research project was that these patients had already received a diagnosis and treatment. This made measures for of demand for (increased GP consultations) difficult to draw from this cohort.

Hypothesis H1<sub>B</sub>: Increased patient self-efficacy is associated with increased demand for healthcare.

TABLE I. RESEARCH SAMPLE OF AGE SPLIT

| Age Split (yrs) | Research sample | DOH sample |
|-----------------|-----------------|------------|
| 16-24           | 2%              | 2%         |
| 25-34           | 4%              | 2%         |
| 35-44           | 7%              | 3%         |
| 45-54           | 9%              | 9%         |
| 55-64           | 25%             | 20%        |
| 65+             | 52%             | 64%        |

This was tested against high versus low demand for healthcare. According to the correlation analysis matrix reported self-efficacy was only linked to age, being negatively correlated ( $p = 0.397$ ). This means that as patients get older their self-efficacy decreases. Whilst there was a good sample size here there was not enough evidence to reject the null hypothesis. The result was therefore not significant.

The key issue with demand for health care is that increasing demand may lead to increase in self-efficacy and this means that patients are less likely to seek health care professional opinions [7]. The findings from this research project suggest that age is negatively correlated with self-efficacy and online information searching; this implies that a younger age group are more likely to search online for health information and take charge of their own health care decisions. This research project concurs with the conclusion in. The author in [3] found in their study that age is negatively correlated with online health information seeking and self-efficacy.

The following hypothesis tests the reported relationship between time on line and demand for healthcare.

Hypothesis: There is a positive relationship between time spent online and demand for healthcare.

This was tested against patients either spending a low amount of time or a high amount of time. Time spent online and demand for health care was supported both by the correlation matrix analysis and by the t test ( $p = 0.011$ ) (one tailed) that was conducted on this research sample. There was evidence to reject the null hypothesis. There was a positive correlation observed between time spent online and demand for health care. Hypothesis H1c also confirmed that there may be a causal link between time spent online and demand for healthcare. This was a significant finding at the 5% level.

## Section 2 Offline and online networks by type

The following hypotheses were used to test the differences between use of two online and two offline forms of information seeking. The following diagram (Fig. 1) maps the two dimensions of on-line and off-line and the amount of active participation. It shows the four possible outcomes from the combination of these two dimensions.

The following represents the hypothesis D, E, F, G, H, and I, the results and the related discussions.

Hypothesis: Patients seeking healthcare information online are more demanding than patients seeking healthcare information offline.

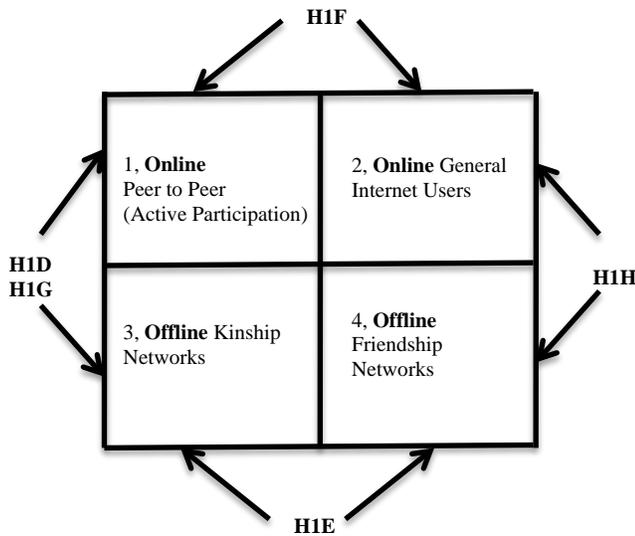


Fig. 1. Offline and Online Networks by Type.

Further exploring demand for healthcare when patients seek healthcare information online using the Chi squared test indicated that there is a statistically significant effect (Chi Squared Test Result:  $p = 0.023$ ) on health care demand compared to patients seeking healthcare information offline. As already discussed the only correlation found with online healthcare information seeking through the correlation matrix analysis was with age where a negative correlation was noted. There was evidence to reject the null hypothesis. This was a significant finding.

This is a significant result and concurs with the results obtained in previous studies [36], [13]. [36] found that searching for healthcare information online has a positive, relatively large and statistically significant effect on demand for healthcare and [13] found that online searches resulted in specific demands for cancer treatments. However this result does not rule out the possibility that patients who seek healthcare information offline are simply less interested in healthcare and those who utilise online methods are more interested.

Hypothesis H1<sub>E</sub>: Increased healthcare information seeking behaviour within online peer to peer patient networks results in increased demand for healthcare, compared with offline 'Kinship' based networks.

The Chi squared test result confirmed that there was no significant difference between online peer to peer networks and offline kinship networks. From this result offline kinship and online peer-to-peer networks cannot be distinguished and have a similar effect on demand for healthcare (Chi Squared Test Result:  $p = 0.487$ ). It is noteworthy that the balance between high and low demand was about equal whereas the split between high and low GP consultation rates in the research sample was 0.62:1 Overall there was not enough evidence to reject the null hypothesis. This was a Non-significant finding.

Hypothesis H1<sub>F</sub>: Patients taking part in offline 'kinship' based networks are more demanding than patients taking part in offline 'friendship' based networks.

The correlation matrix analysis, discussed already, concluded that there was no correlation ( $p = 0.500$ ) between the tested factors with demand for health care. There was no difference observed between off-line kinship and off-line friendship networks. This was a surprise as this test was a repeat of the research done by [31]. It can only be concluded that either the sample size was inadequate to show a statistically significant difference or there is an effect associated with long term chronic conditions which was not evident in the research conducted by [31] and her team. Overall there was not enough evidence to reject the null hypothesis. This was a Non-significant finding.

Hypothesis H1<sub>G</sub>: Patients taking part in online peer to peer networks are more demanding than patients who don't participate in online networks.

The correlation matrix highlighted sharing and engaging online, in other words participating in peer to peer networks, as being positively correlated (Chi Squared Test Result: P value = 0.723) to online information seeking. However a more detailed Chi Squared test did not find a relationship. There was not enough evidence to reject the null hypothesis. This was a Non-significant finding.

Hypothesis H1<sub>H</sub>: The use of non-participatory online networks results in greater demand for healthcare compared with offline friendship based networks.

Chi Squared Test Result: P value = 0.909. The Chi squared test used to compare non-participatory online networks with off-line friendship based networks showed that there was not a statistically significant difference between these two network types. This may be an expected result if there was no 'internet effect' on demand for healthcare when comparing two similar networks whether they be online or offline. A better understanding of the 'internet effect' can be gained by contrasting this result with hypothesis H1<sub>E</sub> which similarly attempted to measure whether there was an 'internet effect' or whether there was a 'social network effect' on demand for healthcare. There was not enough evidence to reject the null hypothesis. This was a Non-significant finding.

Hypothesis H1<sub>I</sub>: The use of social media tools increases demand for specific treatments compared to offline networks.

The Chi squared test was used to establish whether social media tools increase demand for specific treatments. A comparison was made between two cohorts; one that adopted social media tools and one that did not. Whilst there was no significant difference found (Chi Squared Test Result: P value = 0.894), it could be seen that the cohort that adopted online tools within this research project was small for both high and low demand ( $n = 10$  and  $n = 11$  respectively). There was not enough evidence to reject the null hypothesis. This was a Non-significant finding.

Demand for specific treatments as a result of using social media tools represents a significant opportunity for commercial companies. The author in [13] highlighted that this was indeed the case for patients with colon cancer (S. W. Gray, et al., 2009). This research project specifically examined the demand effect for patients with chronic conditions - hypothesis H1<sub>A</sub> above. The measure for demand in this instance, GP consultations, was appropriate as patients who want specific treatments as a result of searching for information online would need to see their clinician. The null hypothesis was accepted. This implies that this patient group are not demanding specific treatments as a result of online searches. The author in [37] in his comparison of online with offline information seeking in older patients and suggests that healthcare decisions based on information obtained 'offline' are more probable in this patient group [37]. Unlike the colon cancer study [13], this patient group was older and perhaps seeking healthcare information offline is more appropriate. This research project explored offline kinship and friendship networks to see if there was a link between offline information searches and healthcare demand (Hypothesis H1<sub>D</sub>), however none was found for this research sample. It is therefore not possible to confirm either [13] or [37] research findings for patients with chronic conditions. This may be that the number of active participants in peer to peer networks in this survey was low as discussed in the limitations of this research.

#### Discussion of findings

Hypothesis H1<sub>D</sub> looked at online versus off-line information seeking and its effect on demand. Based on hypothesis H1<sub>D</sub> this research project supports the idea that utilising the internet compared to offline methods for seeking healthcare information does have a significant effect on demand for healthcare. The result from this research project suggests that some patients seek healthcare information online and in particular younger patients gain a degree of self-efficacy which potentially results in these patients seeking more GP appointments. According to [26], "internet users are more likely to expect they could obtain reliable information about health conditions compared to non internet users" [26]. This result supports the case for a higher degree of motivation amongst online information seekers to seek healthcare information and then do something with that information, for example seeking a clinician appointment. This confirms the results obtained by [36] in that searching for healthcare information online has a positive, relatively large and statistically significant effect on demand for healthcare [36].

Taking Hypothesis H1<sub>E</sub> as a non-significant result; implies that kinship and online peer-to-peer patient communities have no significant difference in the way they act in the context of demand for healthcare. The results from hypothesis H1<sub>E</sub>, non-significant result, when combined with Hypothesis H1<sub>E</sub> infers that there is no 'social network affect' on demand for healthcare, whether that be peer-to-peer or kinship based.

Hypothesis H1<sub>E</sub> was intended to repeat of the work done by [31]. However in comparing the kinship and the friendship networks this hypothesis did not show a statistically significant (off-line) effect.

This again raises the question as to the validity of demand as measured by GP consultation rates. Another explanation is that perhaps patients seeking information online in this research project simply want to explore their current acute condition, a condition that is unrelated to their underlying chronic condition. Overall, however, this research project found no evidence to support peer-to-peer or kinship based influence on demand for healthcare. This concurs with a review [9] which failed to find robust evidence for the benefits of virtual communities about health outcomes [9].

The remaining tests covered by hypotheses H1<sub>F</sub>, H1<sub>G</sub>, and H1<sub>H</sub> resulted in non-significant results. Therefore it is not possible to accept any difference in these methods of information seeking. In particular it is not possible to find a difference between the use of peer to peer networks and general internet searches in increasing demand for health services. This lack of significance is a challenge to the development of the use of peer to peer sites in this context. The following test therefore investigates the demand for specific treatments in the use of social media.

The table (Table II) highlights the summary outcomes of the correlation analysis. What follows is a detailed discussion on the results.

#### Results for Research Objective 1:

To find out which social networks most often used by patients with chronic health conditions.

TABLE II. SUMMARY RESPONSE FROM THE CORRELATION ANALYSIS

| Factor                      | Correlation                              | Pearson's Correlation Coefficient | Strength of Correlation | Sig. (p value) |
|-----------------------------|------------------------------------------|-----------------------------------|-------------------------|----------------|
| Age                         | <i>Self-Efficacy</i>                     | -0.437                            | Moderate                | P = 0.014*     |
|                             | <i>Online Information Seeking</i>        | -0.414                            | Moderate                | P = 0.021*     |
| Online Info Seeking         | <i>Time Spent Online</i>                 | 0.386                             | Moderate                | P = 0.032*     |
|                             | <i>Sharing Health Information Online</i> | 0.404                             | Moderate                | P = 0.024*     |
|                             | <i>Engaging With Others Online</i>       | 0.371                             | Moderate                | P = 0.041*     |
| Offline Info Seeking        | <i>Sharing Health Information Online</i> | -0.556                            | Moderate/Strong         | P = 0.001**    |
|                             | <i>Engaging With Others Online</i>       | -0.523                            | Moderate                | P = 0.002*     |
| Sharing Health Info Online  | <i>Offline Social Networks</i>           | -0.367                            | Moderate/Weak           | P = 0.043*     |
| Engaging With Others Online | <i>Offline Social Networks</i>           | -0.404                            | Moderate                | P = 0.024*     |

From the research sample the respondents could be attributed to each of four categories that are online P2P users, Online general users, Offline Kinship, Offline friendship networks and Offline social networks were easy to identify. However it was difficult to establish whether respondents using ‘offline’ networks ever used the internet for health purposes. Some patients inevitably do and this is reflected by a large ‘online general users’ group.

Referring to section 2.4 and the D.I.D.A. model it was noted that patients who were empowered, ie had a high degree of self-efficacy, may demand more healthcare. According to the D.I.D.A. model if the patient has a positive experience with ‘apomediary’ websites, their knowledge or self-efficacy will increase and the patient will feel more empowered. Consequently their reliance on clinical experts will decrease. Table III highlights websites that respondents in this research project described as ‘favourites’ and therefore have the potential to act as apomediary websites. Of note is that this research uncovered that very few patients (just over 5%) said they regularly visited a ‘favourite’ health website.

Apomediary websites are used to enable patients to become both more self efficacious and autonomous meaning that they depend less on clinicians, known as intermediaries in the Eysenbach’s D.I.D.A. model [7]. The principals of the D.I.D.A. model mean that clinicians today are gatekeepers to treatment whereas previously they were also gatekeepers for information. This information is now being made available online. As internet adoption increases and healthcare information seeking becomes more prevalent so then does the empowering force of the internet on the demand for specific treatments.

Summary Response: Evidence to support adoption of both online and offline networks, however scanty evidence to support adoption of favourite or ‘apomediary’ websites.

TABLE III. LIST OF ‘APOMEDIARY’ WEBSITES

| Website                                                                                                                                         | Description                                    | Number Cited |
|-------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------|--------------|
| <a href="http://www.apparelyzed.com">http://www.apparelyzed.com</a>                                                                             | Spinal cord injury patient website             | 2            |
| <a href="http://hsionline.com">http://hsionline.com</a>                                                                                         | The Health Science Institute                   | 1            |
| <a href="http://www.dbh.nhs.uk/patient-information-leaflets/">http://www.dbh.nhs.uk/patient-information-leaflets/</a>                           | NHS patient information leaflets               | 4            |
| <a href="http://www.mssociety.org.uk">http://www.mssociety.org.uk</a>                                                                           | Multiple Sclerosis Society                     | 3            |
| <a href="https://www.gov.uk/government/organisations/department-of-health">https://www.gov.uk/government/organisations/department-of-health</a> | The Department of Health                       | 1            |
| <a href="http://www.ms-uk.org/newpathways">http://www.ms-uk.org/newpathways</a>                                                                 | Multiple Sclerosis online magazine             | 1            |
| <a href="http://www.nhsdirect.nhs.uk">http://www.nhsdirect.nhs.uk</a>                                                                           | NHS health information                         | 1            |
| <a href="http://www.shinecharity.org.uk">http://www.shinecharity.org.uk</a>                                                                     | Spina Bifida and hydrocephalus support charity | 1            |
| <b>Total</b>                                                                                                                                    |                                                | <b>8</b>     |

Results for Research Objective 2:

To evaluate how patients with chronic health conditions utilise online healthcare information.

It was concluded from hypothesis H1<sub>F</sub> and H1<sub>G</sub> that patients who spend significant time on the internet experience an increase in demand for healthcare. Seeking information online made patients more demanding than patients who seek healthcare information offline. This objective, however, seeks to understand how patients with chronic health conditions use online healthcare information; i.e. does it lead to specific requests for treatment or is it used to comment on other patients’ health situation?

Fig. 2, highlights the extent patients with chronic conditions seek healthcare information online and offline. In this research project there was a 60:40 split between online and offline information seekers in favour of online information seekers.

Patients with chronic conditions, approximately 19%, view other patients’ experiences through online communities, and approximately half of this group will either share their own health experience or comment on others health experiences through these communities. See Fig. 3:

Disclosure of new treatments, such as a possible cure for multiple sclerosis, can drive popularity of social media; it follows that like-minded consumers come together and potentially influence other patients through viral marketing [10]. [4] found that social networking provides ‘social incentives’ meaning that there is a motivation for patients partaking in the sharing of information and engaging with others online [4].

Summary Response: A relatively small but significant proportion of patients view, share and engage online with other patients though patient led communities.

Objective 3

To find out how healthcare companies can better engage with patients who live with chronic health conditions.

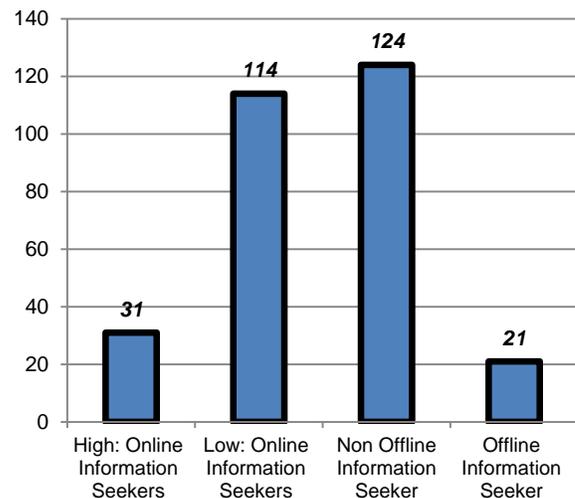


Fig. 2. Online and Offline Information Split by Level.

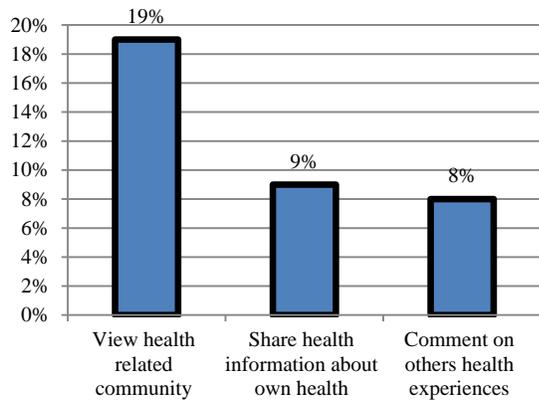


Fig. 3. Viewing, Sharing and Engaging Online.

Companies operating in health care could influence consumer behaviour in several ways based on the findings within this research project. Patients or consumers are potentially most open to influence if they are high internet users, or prefer to take part in peer to peer interactions online with other like-minded patients rather than sharing their symptoms within offline kinship based networks. It is known from this research that patients who use online search methods demand more healthcare than patients who adopt offline methods. It is also known that the more time a patient spends online the more demanding the patient becomes. Patient forums are popular with some patients and indeed these forums or communities encourage patients to spend more time online. Whilst this research project failed to show a correlation with peer to peer networks and demand for healthcare; a relationship between time spent online and P2P network usage was shown to exist, see Table IV.

Patients using P2P communities spend more time online. This may be a consequence, albeit not proven in this research project, that patients who use these networks may be more demanding of healthcare as a result of the time spent online. If healthcare companies encourage patients to utilise P2P communities via their own websites then this may encourage increased time spent online and consequently encourage an increase in demand for healthcare e.g. demanding specific treatments.

TABLE IV. PEER TO PEER USAGE ACCORDING TO TIME SPENT ONLINE

| Actual Result            | P2P non user | P2P User | Total | Exp. Result              | P2P non user | P2P User | Total |
|--------------------------|--------------|----------|-------|--------------------------|--------------|----------|-------|
| Time Spent Online = High | 11           | 9        | 20    | Time Spent Online = High | 16           | 4        | 20    |
| Time Spent online = low  | 101          | 18       | 119   | Time Spent online = low  | 96           | 23       | 119   |
| Total                    | 112          | 27       | 139   | Total                    | 112          | 27       | 139   |

Chi Squared Test Result: P value = 0.0018\*

\* significant at 1% level

Summary Response: Evidence to support the link between P2P network usage and time spent online but insufficient causal evidence to support P2P network usage and increase demand for healthcare.

Objective 4

To determine demographic differences in adoption of online social media tools used by patients with chronic health conditions.

TABLE V. GENDER SPECIFIC INTERNET ADOPTION RATES

| Gender | High Internet Adoption (n) | Low Internet Adoption (n) | Percentage High Adoption |
|--------|----------------------------|---------------------------|--------------------------|
| Female | 12                         | 43                        | 22%                      |
| Male   | 19                         | 68                        | 22%                      |
| Total  | 31                         | 111                       | 22%                      |

The table (Table V) highlights the fact that there were no gender differences in internet adoption rates in this research sample. The author in [15] found that females were more likely to use the internet for health related information than males and were more likely to belong to patient support groups [15]. This effect was not observed in this research project, possibly a reflection of the smaller size of the female segment.

TABLE VI. GENDER SPECIFIC SELF EFFICACY I.E SEEKING DISEASE AND TREATMENT INFORMATION

| Statistical Comparison     | Variable               | Comparison | Sample Size | Mean | Variance | P value one tailed |
|----------------------------|------------------------|------------|-------------|------|----------|--------------------|
| t Test independent samples | Gender & Self Efficacy | Female     | n = 55      | 3.51 | 0.82     | p = 0.063          |
|                            |                        | Male       | n = 87      | 3.10 | 1.26     |                    |

This table (Table VI) illustrates gender specific differences in self-efficacy, known to be a key factor in demanding more healthcare information and potentially asking the GP for specific treatments. Statistically there are no gender specific effects for self-efficacy.

Summary Response: There was evidence to support increased internet adoption rates according to defined age bands, indicating that being unwell increases likelihood of internet use compared to the general population. There were no gender differences noted.

Research Question

Have online social media tools affected demand for healthcare in patients who experience chronic medical conditions?

This research project provides evidence to support the idea that searching for healthcare information and spending more time online has positive effects on demand for healthcare amongst patients who live with chronic medical conditions. Patients who have chronic conditions spend more time online than the general population. Patients who take part in online peer to peer communities are more trusting of healthcare

information obtained in an online setting. There was a causal link between time spent online and Peer to Peer network adoption. Spending time online was shown to have a positive effect on demand for healthcare. This research project also highlighted that younger patients are likely to become more self-efficacious and this potentially translates into a greater autonomy relating to their healthcare treatment decisions. In future these younger patients will age and replace the older generations with their social networking preferences. These patients will spend increasing time on peer to peer communities and as a consequence their demand for healthcare in general for treatments specifically will increase. The patients of the future will therefore depend less on their doctor for healthcare or product information and see the doctor as simply a gatekeeper for accessing treatments. The doctors of the future will experience patients who are more demanding, not necessarily for health information but for specific treatments.

TABLE VII. OVERVIEW OF HYPOTHESIS, TEST TYPE, P VALUES AND OUTCOME

| Hyp             | Detail                                                                                                                                                                                                          | Statistical Comparison     | P value   | Null Hypothesis Accepted or Rejected |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------|-----------|--------------------------------------|
| H1 <sub>A</sub> | Age has a positive effect on health information seeking behaviour online                                                                                                                                        | t Test independent samples | p = 0.068 | Accepted                             |
| H1 <sub>B</sub> | Increased patient Self Efficacy is associated with increased demand for healthcare                                                                                                                              | Linear Regression          | p = 0.397 | Accepted                             |
| H1 <sub>C</sub> | Patients taking part in increased healthcare information seeking behaviour within online peer to peer patient networks seek more GP consultations compared with patients using offline 'Kinship' based networks | Chi Squared Test           | p = 0.487 | Accepted                             |
| H1 <sub>D</sub> | Patients taking part in offline 'kinship' based networks are more demanding than patients taking part in offline 'friendship' based networks                                                                    | Chi Squared Test           | p = 0.500 | Accepted                             |
| H1 <sub>E</sub> | Patients taking part in online peer to peer networks are more demanding than patients who don't participate in online networks                                                                                  | Chi Squared Test           | p = 0.728 | Accepted                             |

|                 |                                                                                                                                       |                            |            |          |
|-----------------|---------------------------------------------------------------------------------------------------------------------------------------|----------------------------|------------|----------|
| H1 <sub>F</sub> | There is a positive relationship between time spent online and demand for Healthcare                                                  | t Test independent samples | p = 0.011* | Rejected |
| H1 <sub>G</sub> | Patients seeking healthcare information online are more demanding than patients seeking healthcare information offline                | Chi Squared Test           | p = 0.023* | Rejected |
| H1 <sub>H</sub> | The use of non-participatory online networks results in greater demand for healthcare compared with offline friendship based networks | Chi Squared Test           | p = 0.909  | Accepted |
| H1 <sub>I</sub> | The use of social media tools increases demand for specific treatments compared to offline networks                                   | Chi Squared Test           | p = 0.894  | Accepted |

\* significance at 5% level

The table (Table VII) highlights the hypotheses tested, the test, result and the outcome pertaining to the acceptance or rejection of the hypotheses.

#### IV. LIMITATIONS OF RESEARCH

The sample targeted represented patients that all had chronic long term conditions. This was identified by them being on treatment for a significant length of time; unfortunately this also meant that these patients were not recently started on treatment. Therefore the possibility exists that these patients might have referred to health conditions that were not related to their primary diagnosis when completing this questionnaire. This could have affected their interpretation of the questions which would have a negative effect on the validity of the results obtained.

The literature clearly suggested that there were differences in demand noticed as a result of the use of social media tools. However there remains the possibility that patients may perceive to have their demand for healthcare satisfied through their online search for healthcare information. Thus resulting; however, in a decrease in essential GP consultation, a key outcome of demand. However there is a significant likelihood that misinformation and mis-diagnosis can result in further complications due to the reduction of GP consultations.

For companies in healthcare that currently have no social media policy, any patients registered on the company's Direct To Patient service, would find very limited information on products through social media. In contrast other manufacturers in the healthcare sector that do have a social media policy may be at an advantage. Potentially this creates a bias in so much that patients who are familiar with social media tools may 'prefer' products from manufacturers who have a social media policy. This introduces a potential bias in this study whereby

the testing of patients from other healthcare Direct To Patient Services with social media provision was not undertaken.

## V. CONCLUSION

1) *The changing doctor-patient relationship:* From this research paper the link between age and sharing and engaging online was unproven. Age was however negatively correlated with online information seeking behaviour and self-efficacy. This implies that the younger patient is taking more control of their health and seeking more information online, a trend that is likely to increase as these young patients become old patients.

2) *The patient as the new decision maker:* More than ever before it is possible to listen to the consumers' voice through either producer or community led communities and acting on their feedback. Involving the patient in the product or healthcare decision making process serves to not only add value but can improve healthcare outcomes especially patient compliance. In an attempt to overcome poor compliance it has been suggested that patient involvement through the concordance model would achieve greater self-efficacy [39]. A concordance between the expert and the patient requires the mutual adoption of the treatment regime. This mutuality often requires a positive efficacious embrace of the adopted solution. The consequence of this is that patients often are now much more demanding of their healthcare professional; appointment times are however increasing due to Covid and demands upon healthcare professionals are ever more stretched meaning that the burden of responsibility for patient knowledge and healthcare choices falls increasingly upon the patient [3]. The key here being that the patient can contribute more, regarding symptoms that can better inform the GP to enable more accurate diagnosis. As already discussed self-efficacy is negatively correlated with age but was not shown to be related to demand for healthcare in this research project. The issue of measuring demand was discussed and a further understanding of outcomes expected as a result of online searches is needed.

3) *A model for seeking Health information:* Eysenbach's D.I.D.A. model explored the idea of relying on apomediarities [7]. These may be specific 'one stop shop' websites or peer to peer networks. It was apparent from this research project that patients taking part in patient peer to peer communities either online or offline were no more demanding in terms of GP consultation requests. There was no 'network effect' observed offline which contradicted [31] work. It can be concluded that modern day healthcare information seekers behave differently to the kinship of friendship networks observed by [31].

4) *Healthcare information seeking increases demand for health care:* Spending more time online, from this research project, was shown to have a positive effect on demand for healthcare. It was also demonstrated that patients seeking healthcare information online were more demanding than patients who seek healthcare information offline, meaning that patients who spend time online and seek healthcare

information are likely to seek more GP consultations. However whilst these patients are consulting more with their GP they are not, according to this research, demanding specific treatments.

## Recommendations

Concerning the patient population in this study; patients who experience ongoing changes in their treatments as a result of their condition (e.g. MS patients - whose condition evolves over time) may have been a better group of respondents to judge the effectiveness of social media and its influence. Given the difficulties in determining demand for healthcare inherent in this research project, a re-examination of the research philosophy is suggested. The chosen philosophy was positivism as this emphasised the value of predicting outcomes of the research so that these variables might be controlled in future. According to [36] at the root of the positivist research philosophy is the law of cause and effect [36].

Examples of cause and effect in this research project are:

- 1) Social networks 'cause' an increase in healthcare information seeking behaviour - the 'effect'
- 2) Age 'causes' increase in healthcare information seeking behaviour - the 'effect'
- 3) Spending time online 'causes' an increase in demand for healthcare - the 'effect'

Essentially, and as discussed previously, assumptions were made about what causes demand and what 'demand' is in the context of healthcare. Therefore a better way of determining the factors or 'causes' that are likely to predict the outcome, the 'effect' is needed. According to [36] greater organisational or online market complexity, with the online/offline environment, would lead towards an interpretivist approach [36]. Referring to interpretivism makes it necessary to conduct research in the online environment in order to understand what is going on among the markets 'social actors'; people who play a part on the stage of online and offline information seeking. Building on the correlation analysis the demand for healthcare could be further explored using an interpretivist approach to better understand what patients with chronic conditions are looking for.

This research project represents a snapshot in time of people's current internet adoption rates. Therefore as age and usage of technology increases in future, the age correlation is expected to become less negative meaning that patients across all levels will have greater self-efficacy in their online searching behaviours. This means that the younger patient with a chronic condition will, if they stay with the company's Healthcare's DTP service, become the next older patient generation and as such be more accustomed to using online search tools.

The author in [23] suggests that a stronger bond between producer led and customer (the patient) led community interactions will better enable both an understanding and adaptation of the marketing message. The findings from this research project suggest that peer to peer communities are more trusting of online information and spend more time

online; typically these patient groups are patients with chronic conditions. These patient led communities represent an opportunity to augment the traditional direct marketing interface which represents a producer: customer interaction. This may take the form of directing patients to the company's own website which may not contain traditional marketing or product related information but rather contains the resources required for the patient to begin to see the company website as a 'one stop shop' or favourite website, known as an apomediary website. This would provide the patient with the patient networks and information needed to help the patient form positive impressions of the website, and the company that produced it. Ultimately the patient will express a preference for the company's products once they visit their GP.

## VI. FUTURE RESEARCH

From this research project it is clear that more work would need to be done to corroborate the motivation patients have for seeking healthcare information online and how this translates into more demand for healthcare. Thus:

Establishing the motives for healthcare information seeking through a qualitative analysis and structured interviews would prove insightful.

Exploring the extent to which patients with chronic conditions depend on the internet and social media tools would also prove invaluable.

This research project looked at patients who had already received their primary diagnosis. What happens to patients before that diagnosis? Is it possible that patients who were prolific internet users prior to receiving their initial diagnosis and once established on treatment their internet usage waned? This research project was unable to establish internet and social media tool usage prior to the patients receiving treatments.

It is well known that an established friend on Facebook follows an 'offline to online' trend with people making friends offline first and then later adding them online [27]. Does the same 'offline to online' affect exist with patients who have long term conditions such as spinal cord injury? For example do these patients later connect online with other patients met 'offline' within spinal cord units?

According to [22] Social psychologists make a distinction between different attitude levels with the deepest being referred to as personality, followed by values then attitudes with the most superficial being referred to as opinions [22]; establishing the effect of the patients' personality on social media usage, for example the degree of extroversion or introversion. Could this be a motivation to experiment with seeking online health information rather than utilizing established offline methods? Conversely introversion is associated with a greater dependence on Facebook for communicating with and establishing friends (Orr et al.,2009); could the same influence exist with patients with chronic conditions?

It would be valuable to establish a long term view of how patients with chronic conditions use social media tools post

Covid, particularly as this study was pre-Covid. This research project took a cross sectional view on social media adoption. A longitudinal study could follow the patient's use of social media tools from before diagnosis to a time period where the patient was established in terms of both their condition and treatment. This would give invaluable insights into the patient's use of social media tools.

This research project only looked at patients with long term conditions; it would also be helpful to consider patients with short term acute conditions to establish differences in motivations for seeking healthcare information and support.

An underlying tenet of this study was that use of social media tools affects demand for healthcare. It was assumed that demand could only be in the form of seeking clinician appointments or specifically asking for treatments. Additional needs affecting demand may also exist particularly with patients using peer to peer networks.

## REFERENCES

- [1] Al-Qahtani, M.F., Al-Saffar, A.K., Alshammasi, A., Alsanni, G., Alyousef, Z.T., & Alhussaini, M. (2018). Social media in healthcare: Advantages and challenges perceived by patients at a teaching hospital in eastern province, Saudi Arabia. *Saudi Journal for Health Sciences*, 7, 116 - 120.
- [2] Benetoli, A., Chen, T. F., & Aslani, P. (2017). Consumer Health-Related Activities on Social Media: Exploratory Study. *Journal of medical Internet research*, 19(10), e352. <https://doi.org/10.2196/jmir.7656>
- [3] Cotton, S. R., & Gupta, S. S. (2004). Characteristics of online and offline health information seekers and factors that discriminate between them. *Social science & medicine*, 59(9).
- [4] Crosier, B. S., Webster, G. D., & Dillon, H. M. (2012). Wired to connect: Evolutionary psychology and social networks. *Review of General Psychology*, 16(2).
- [5] Davis, T. F. (2012). What marketers say about working online. *McKinsey Quarterly*.
- [6] Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, 7(1).
- [7] Eysenbach, G. (2008). Medicine 2.0: social networking, collaboration, participation, apomediation, and openness. *Journal of Medical Internet Research*, 10(3).
- [8] Eysenbach, G., & Kohler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*, 324(7337).
- [9] Eysenbach, G., Powell, J., Englesakis, M., Rizo, C., & Stern, A. (2004). Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*, 328(7449).
- [10] Gabarron, S.E., Fernandez-Luque, L., & Armayones, M. (2012). Social media in health -- what are the safety concerns for health consumers? *Health Information Management Journal*, 41(2), 30-35.
- [11] Girard, T., Dion P., (2010), Validating the search, experience, and credence product classification framework. *Journal of Business Research*, Vol.63, pp1079-1087.
- [12] Gray, N. J., Klein, J. D., Noyce, P. R., Sesselberg, T. S., & Cantrill, J. A. (2005). Health information-seeking behaviour in adolescence: the place of the internet. *Social science & medicine*, 60(7), 1467-1478
- [13] Gray, S. W., Armstrong, K., DeMichele, A., Schwartz, J. S., & Hornik, R. C. (2009). Colon cancer patient information seeking and the adoption of targeted therapy for on-label and off-label indications. *Cancer*, 115(7).
- [14] Hagg, E., Dahinten, V. S., & Currie, L. M. (2018). The emerging use of social media for health-related purposes in low and middle-income countries: A scoping review. *International journal of medical informatics*, 115, 92-105. <https://doi.org/10.1016/j.ijmedinf.2018.04.010>

- [15] Hesse, B. W., Nelson, D. E., Kreps, G. L., Croyle, R. T., Arora, N. K., Rimer, B. K., et al. (2005). Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. *Archives of Internal Medicine*, 165(22), 2618.
- [16] Jacques Bughin, A. H. B., Michael Chui. (2011). How social technologies are extending the Organisation. *Mcknsey Quarterly* (November), 1-10.
- [17] Karla Anderson, L. S., Garrett, D., (2012). Social media "likes" healthcare for marketing to social business: Health research Institute.
- [18] Lanther, E. (2002). *Psychology Research Methods*. Retrieved 11/11/2013, from <http://www.nvcc.edu/home/elanthier/methods/index.htm>
- [19] Murray, E., Lo, B., Pollack, L., Donelan, K., Catania, J., White, M., et al. (2003). The impact of health information on the internet on the physician-patient relationship: patient perceptions. *Archives of Internal Medicine*, 163(14), 1727.
- [20] Ofcom. (2012). *The Communications Market 2012*: Ofcom.
- [21] Office for National Statistics (2012), <http://www.ons.gov.uk/ons/rel/rdit2/internet-access--households-and-individuals/2012/stb-internet-access--households-and-individuals--2012.html>
- [22] Oppenheim, A. N. (2003). Questionnaire design, interviewing, and attitude measurement: Pinter Pub Ltd.
- [23] Palmer, A., & Koenig-Lewis, N. (2009). An experiential, social network-based approach to direct marketing. *Direct Marketing*, 3(3), 162-176.
- [24] Parker, R., & Pettijohn, C. E. (2003). Ethical Considerations in the Use of Direct-To-Consumer Advertising and Pharmaceutical Promotions: The Impact on Pharmaceutical Sales and Physicians. *Journal of Business Ethics*, 48(3), 279-290.
- [25] Powell, J., & Clarke, A. (2002). The WWW of the World Wide Web: who, what, and why? *Journal of Medical Internet Research*, 4(1).
- [26] Rice, R. E. (2006). Influences, usage, and outcomes of Internet health information searching: multivariate results from the Pew surveys. *International Journal of Medical Informatics*, 75(1).
- [27] Ross, C., Orr, E. S., Sisic, M., Arseneault, J. M., Simmering, M. G., & Orr, R. R. (2009). Personality and motivations associated with Facebook use. *Computers in Human Behavior*, 25(2), 578-586.
- [28] Roxane Divol, D. E., Hugo Sarrazin (2012). *Demystifying Social Media*. [Marketing and Sales Practice]. McKinsey & Company (April), 1-11.
- [29] Saren, M. (2011). Marketing empowerment and exclusion in the information age. *Marketing Intelligence & Planning*, 29(1), 39-48.
- [30] Saunders, M. (2012). *Doing Research in Business & Management An Essential Guide to Planning Your Project*: Prentice Hall.
- [31] Scambler, A., Scambler, G., & Craig, D. (1981). Kinship and friendship networks and women's demand for primary care. *The Journal of the Royal College of General Practitioners*, 31(233), 746.
- [32] Sheth, J. N., & Parvatiyar, A. (1995). Relationship marketing in consumer markets: antecedents and consequences. *Journal of the Academy of marketing Science*, 23(4).
- [33] Sontag, S. (2001). *Illness as metaphor and AIDS and its metaphors*: Picador.
- [34] Number of social media users worldwide 2010-2021 [Internet]. [cited 2020 Mar 27]. Available from: Statista. [www.statista.com](http://www.statista.com) Internet Social Media & User-Generated Content
- [35] Susannah Fox, M. D. (2013). *Health Online Pew Internet & American Life Project*.
- [36] Suziedelyte, A., (2012). How does searching for health information on the Internet affect individuals' demand for health care services?, *Social Science & Medicine*, Vol 75, Issue 10, pp 1828-
- [37] Taha, J., Sharit, J., & Czaja, S. (2009). Use of and satisfaction with sources of health information among older Internet users and nonusers. *The Gerontologist*, 49(5), 663-673.
- [38] Ventola, C. L. (2008). Challenges in evaluating and standardizing medical devices in health care facilities. *Pharmacy and Therapeutics*, 33(6), 348.
- [39] Vermeire, E., Hearnshaw, H., Van Royen, P., & Denekens, J. (2002). Patient adherence to treatment: three decades of research: A comprehensive review. *Journal of clinical pharmacy and therapeutics*, 26(5), 331-342.
- [40] Xiao, N., Sharman, R., Rao, H.R., Upadhyaya, S., (2014), Factors influencing online health information search: An empirical analysis of a national cancer-related survey, *Decision Support Systems*, Vol 57, pp 417-427
- [41] Ybarra, M., & Suman, M. (2008). Reasons, assessments and actions taken: sex and age differences in uses of Internet health information. *Health Education Research*, 23(3).
- [42] Zhao, Y., & Zhang, J. (2017). Consumer health information seeking in social media: a literature review. *Health information and libraries journal*, 34(4), 268-283. <https://doi.org/10.1111/hir.12192>
- [43] Zickuhr K, A. S. (2012). *Digital differences: Pew Internet & American Life Project*.

# Predicting Cyber-Attack using Cyber Situational Awareness: The Case of Independent Power Producers (IPPs)

Akwetey Henry Matey<sup>1</sup>, Paul Danquah<sup>2</sup>, Godfred Yaw Koi-Akrofi<sup>3</sup>

Department of I.T. Studies, University of Professional Studies, Accra (UPSA), Ghana<sup>1,3</sup>  
Department of I.T. Heritage Christian College, Accra, Ghana<sup>2</sup>

**Abstract**—The increasing critical dependencies on Internet-of-Things (IoT) have raised security concerns; its application on the critical infrastructures (CIs) for power generation has come under massive cyber-attack over the years. Prior research efforts to understand cybersecurity from Cyber Situational Awareness (CSA) perspective fail to critically consider the various Cyber Situational Awareness (CSA) security vulnerabilities from a human behavioural perspective in line with the CI. This study evaluates CSA elements to predict cyber-attacks in the power generation sector. Data for this research article was collected from IPPs using the survey method. The analysis method was employed through Partial Least Squares Structural Equation Modeling (PLS-SEM) to assess the proposed model. The results revealed negative effects on people and cyber-attack, but significant in predicting cyber-attacks. The study also indicated that information handling is significant and positively influences cyber-attack. The study also reveals no mediation effect between the association of People and Attack and Information and Attack. It could result from an effective cyber security control implemented by the IPPs. Finally, the study also shows no sign of network infrastructure cyber-attack predictions. The reasons could be because managers of IPPs had adequate access policies and security measures in place.

**Keywords**—Internet of things; cyber situational awareness; critical infrastructures; power generation; cyber-attack; cyber security; human behavioural and independent power producers

## I. INTRODUCTION

The massive application of IoT on the electrical grid opened up a huge opportunity to utilize previously untapped processing power to offload custom applications directly to other devices. These transformations are not achievable without experiencing some form of security vulnerabilities on the grid ([1], [2]), even though deploying these business applications on the grid will increase the overall robustness of the grid and reduce communication overhead. A recent attack in 2016 on the Ukrainian power grid was an advanced form of hacking the CI by the Russians extending their intrusion to increase control using the “Crash Override”. A related attack occurred in March 2016, compromising the command-and-control (C&C) system on the New York Dam with a cellular phone. The Stuxnet attack also created awareness of potential cyber threats on power generation companies[3], impacting the grid's reliability [4]. According to [5], various

forms of cyber-attack concerning the grid are man-made manipulation. Raikuma, et al [6], have indicated the need to avoid such incidents due to the ripple effects of a power system shutdown. The growing investment of human capital and financial resources injected into CI protection shows the extent to which industry players and the research community understand CI challenges. The increased investment in the sector calls for the need to evaluate cyber situational awareness (CSA) from a human behaviour perspective since the critical infrastructure (CI) falls within a dynamic changing environment. CSA can assist in comprehensively investigating an approach to the ongoing debates relating to cyber security. Cyber awareness for cyber defence generally requires perception, understanding and projection. CSA creates room for predictions in line with an action sequence and effectively plans for new cyber-attacks trends. Hence, the need to identify activities of interest to maintain awareness of a new paradigm in cyber defence. According to Franke and Brynielsson [7], CSA is comparable to insider informants leaking information on an imminent attack. Prior studies had also revealed a range of cyber incidents stemming from minor employee mistakes, misinformation on controls, and highly coordinated, well-planned attack on the critical infrastructure ([1], [8]–[10]). Johnson and Banfield [11] revealed that current cybersecurity defences cannot match the sophistication of embedded technologies attacks capabilities. Hence, the need to evaluate cyber situational awareness (CSA) from human behaviours perspective to predict, detect, and prevent cyber-attack vulnerabilities in a dynamic power generation environment. The concept of cyber situational awareness can be situated based on the insight of individual abilities to distinguish and assess current and future effects in terms of how situations evolve from an attacker's perspective to understand and restrict cyber vulnerabilities [12]. Prior research has made an immense contribution in applying various technologies to support critical cyber incidents. Yet, today's cybersecurity challenges in the power generation sector are increasing and becoming more sophisticated and alarming than we think in the power sector ([5], [13], [14]). The introduction of intelligent information technology equipment such as the Internet of things (IoT) devices and other industrial control systems (ICS) enabled the power generation grid to become more effective and intelligent ([3], [15], [16]). Fig. 1 depicts a visual overview of a possible cyber-attack on the generation and distribution section of the power grid.

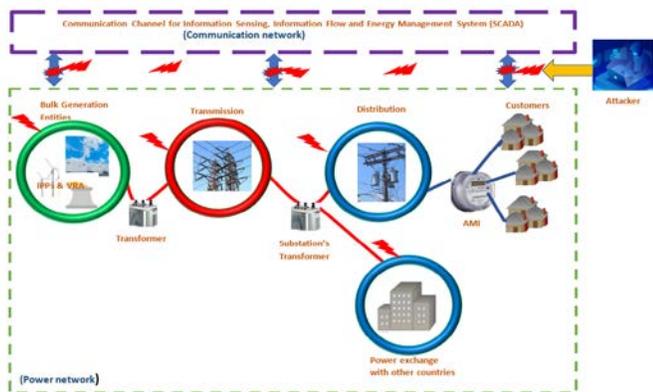


Fig. 1. Smart Grid Perspective in Ghana Source: Authors' Construction (2021).

Based on the above discussion, this current study seeks to establish how CSA elements (Network, Information and People) influence cyber-attacks in the power generation environment in perspective. Digital trust links people, data, and networks [17]. The specific objectives of this study are therefore to:

- 1) Evaluate how staff perceives potential cyber threats in their working environment.
- 2) Evaluate information vulnerabilities and how they influence cyber Attack.
- 3) Assess how cyber situational awareness network vulnerabilities contribute to cyber Attack.

## II. LITERATURE REVIEW

### A. Theoretical Review

Endsley [18] gave three (3) levels of indication to assist in forming a mental model of having a more comprehensive view of an operational environment, as shown in Fig. 2.

In our effort to evaluate cyber situational awareness from human behaviour in an IPPs environment, the authors consider the base of the variable in the conceptual framework in line with the considerations in Table I and Fig. 3.

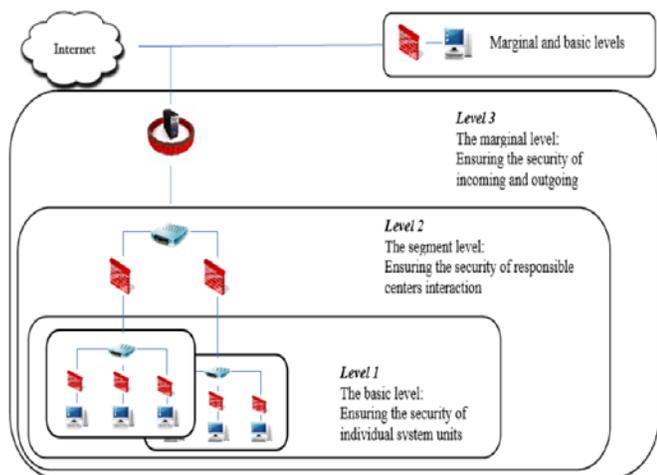


Fig. 2. A Model for Understanding Operational Situation by Endsley.

TABLE I. SHOWS THE MAP OF AUTHORS CONSIDERATIONS AND ENDSLEY

| Focus | Endsley                                   | Authors Considerations                                                                                                  |
|-------|-------------------------------------------|-------------------------------------------------------------------------------------------------------------------------|
| 1     | Security of individual systems            | Security from human behavioural in context interacting with the system of the IPPs                                      |
| 2     | Security of Centre's                      | Security from the perspective of the IPPs Grid Network                                                                  |
| 3     | Security concerning incoming and outgoing | protection from human behavioural in context when dealing with sensitive and confidential pieces of information of IPPs |



Fig. 3. A Proposed Conceptual Framework Source: Authors' Construction (2021).

The basis for each of the hypotheses used in this study is explained below.

### B. People

In 2019 the Worldwide Threat Assessment by the U.S. indicated that hackers, hacktivists, and insiders pose significant cyber threats to the grid. According to [19], a variation of threat actors pose substantial cybersecurity threats to the electric grid; these actors support grid operations. Ramamurthy and Jain [16] also indicated the difficulty in managing the workforce transformation, which is likely to be the most worrying complication of IoT implementations. Cyberattacks at varying levels of criticality on nations businesses and organizations with an internet presence primarily contribute to human-centric activities [20]. Some persons may work as a team to make decisions and carry out actions [18]. In the work of [21], internal employees working on the grid due to untrained employees or unhappy employees who have hatefulness for other consumers or the service providers also contribute to cyber-attack activities, therefore, seeking to evaluate the vulnerabilities. These factors are considered based on social networking, operating procedure maintenance, and security-related issues, such as user elicitation of information on cyber threats on CI networks, person-to-person interactions challenges concerning user control, and whether or not users strictly adhere to cyber security protocols. Based on the preceding, we propose the following hypothesis.

(H1): Operational staff's cyber security vulnerabilities positively influence cyber-attacks.

(H1.1): Vulnerabilities from operational staff effect of network activities positively influence cyber attacks.

### C. Information

In conceptualizing cyber situational awareness, there is the need to stress the practical concept that conveys security-

relevant information that supports the decision-making process [7]. In context, we refer to the logical data flow between network nodes, such as the IoT devices and other intermediary devices, which mandate is to temporarily collect and transmit some form of data emanating from CI activities. A recent report [19] indicated that a critical protection assessment action required to address cybersecurity risks facing the electric grid is Data security and Information protection processes and procedures. Due to power generation system vulnerabilities to cyberattacks against practical state estimation, Zhang [22] develops a comprehensive situational awareness framework for distribution system information on monitoring and controlling state estimation components, cyber-attack detection information, fault location, and voltage control information. Frequently, receiving devices generate information obtained from different devices sources and determine its reliability. Nonconformities between the essential information and possible vulnerabilities areas in the grid can be recognized from emerging information's nature [23]. The constructs objective is to evaluate vulnerabilities and their ramifications on the grid.

- The authors measured how individuals apply cyber security controls when dealing with sensitive and confidential information.
- Whether they had ever experienced information leakages or sensitive information received from the grid comes with inaccuracies.
- If there are restrictions on remote access and finally to enquire if there had been pieces of evidence of frequent misleading information been to receive on their systems.

(H2): The Information handling from operational staff positively influences cyber-attack.

(H2.1): The effect of network activities on information handling influences the cyber Attack.

#### D. Network

The communication network and the electrical grid play a significant role in generating and transmitting power to either a sub-station or a customer. Hence, identifying various network vulnerabilities impact assessments will provide knowledge of future impact projection. Prior studies have discussed different techniques to improve cyber situational awareness [7], [24], [25], mainly for analyzing the trends in network traffic. With the evolution of grid networks, there are increasing security threats due to the expanding volume of data transmitted on the grid. Some of these specific cyber incidents on the grid network, as indicated by ([26]–[28]) in their recent study. Because Cyber situation awareness empowers cybersecurity experts to detect and fully understand and anticipate incoming threats, however, our thorough review of the literature revealed no previous studies on how the human behavioural concept can apply in this context. To assess cyber situational awareness of the grid network vulnerabilities and how these vulnerabilities can influence cyber-attacks. Per these constructs, our goal is to evaluate

vulnerabilities and their implications for the grid. Concerning the measurement of network vulnerabilities from a human behavioural perspective, authors seek to:

- Enquire whether or not individuals can access the grid network with their devices.
- Find evidence of unauthorized persons accessing the grid network remotely.
- Find the frequency of change of network access policies.
- Enquire if there is evidence of unauthorized IT staff accessing the network remotely.
- Finally, enquire if users can access social media applications on the network.

H3. Negative human behavioural activities on the network positively influence cyber-attacks.

#### E. Cyber Situational Awareness

The military is where situational awareness first appears. Situational awareness aims to identify events, causes, consequences, and future projections [18]. It also considers the status and attributes of elements by assessing the present situation and predicting future outcomes based on previous understanding and acquaintance [29]. It becomes possible through the acquisition of data, conception, and synthesis to enable decision-makers to resolve problems with the massive deployment of IoT devices in the power generation sector; for data, acquisition to continuously monitor various sub-systems of the entire power generation system Infrastructure. Therefore, Cyber Situation Awareness (CSA) extends Situation Awareness (SA) to the power generation cyber domain. Hence, we can access fist hand information and seek indications from an attacker's perspective, estimate the impact, to anticipate their actions. Research has carefully refined cyber situational awareness predominantly in the CI ([22], [30], [31]). However, a careful study of the literature did not reveal any prior studies investigating how the Cyber Situation Awareness elements (people, information, and networks) concept can apply in human behaviours from IPPs operational environment. Because according to Michael, et al [32], CSA is the degree to which individuals within a team possess the CSA required in carrying out their responsibilities. We believe cyber vulnerabilities could occur due to various duties discharged by the operational staff of IPPs hence, evaluating human behaviours in the power generation sector. These vulnerabilities can occur in any of the layers; physical layer, information layer, and the human layers in the CI [33]. Also see Appendix.

#### F. Cyber-Attack

Cyber-attack issues relating to the smart grid and its impact on the IPPs are increasingly problematic and threatening to a developed and developing economy. In terms of business and human privacy and even national security, the current grid Infrastructure uncertainties manifest due to the power generation sector's Internet of things (IoT). Its cyber security issues have become a significant subject of debate

globally. Krishnan, et al [34] indicate evidence of price cap and bid price manipulation and subsequent Attack on the generation unit. Memories of a cyber attack in New York in March 2016 affected the Dam control system with a cellular modem. Such attacks can lead to incapacitating the practical function of the electric grid in line with communication between systems or equipment on the grid network ([8], [15], [35]). It can also harm the effective grid functioning ([1], [36], [37]). Recent studies in the area reveal various cyber-attacks such as False data injection attacks ([27], [38], [39]), Denial-of-service attacks identified by ([40], [41]), Distributed Denial of Service attack indicated by [42], Man-in-the-middle attack [6], Malware Attack by ([43], [44]), State Estimation attack [45], and Price manipulation and Misrepresentation of values attack[1] Coordinated ([10], [43]), etc.

### G. Related Work

The new paradigm of human-centric warfare on cybersecurity is wealth looking into because insider threats can be disruptive and equally malicious as an attack from outside an organization [20]. Cyber situational awareness gives insight into understanding an impending phenomenon [46]. Since many sectors in the economy are primarily dependent on Critical infrastructures (CI) [11], where information is sent and received for processing to predict possible future threats and adequately plan the power generation environment. According to Nekha and Dorosh [23], the critical aspect of cyber situational awareness can timely deal with an emerging threat model. Hence, critically examine elements such as people, information, and networks in the context of cyber situational awareness within the operational environment of the independent power producers (IPPs) in perspective. The growing adoption of high IoT devices connected to the Internet and the use of the global positioning system to harmonize grid operations contributes to grid vulnerabilities [19]. Current studies in the sectors focus on the cyber-physical aspect of the grid; [47] examine the multi-stage attacks using WannaCry ransomware. The study of [48] accesses the vulnerability with the local power trading. Sharafeev et al [1] develop an algorithm to monitor electrical power systems (EPS) cyber-attacks. A framework for assessing critical infrastructure in an attack was created by Akhtar, et al [49]. Sarangan, et al [50] Analyzed cyber-attacks of the power grid in considering the ill-effects of increasing renewable penetration [39]. Develop a systematic two-stage approach for detecting false data injection (FDI). Roy and Debbarma [27] also Proposed a cyber-attack detection and mitigation platform which uses forecasted data. Raikuma, et al [6] also demonstrate the impact of the man-in-the-middle attack, which exploits vulnerabilities in the Generic Object-Oriented Substation Event (GOOSE). Prior related research work was sort of giving indications of lack of effort to explore vulnerabilities from Human Behavioral perspective particularly, in the power generation sector. Hence our current study seeks to critically evaluate and access the cyber vulnerabilities within the IPPs using cyber situational awareness.

Based on the above reviews, we propose five (5) hypotheses, as follows:

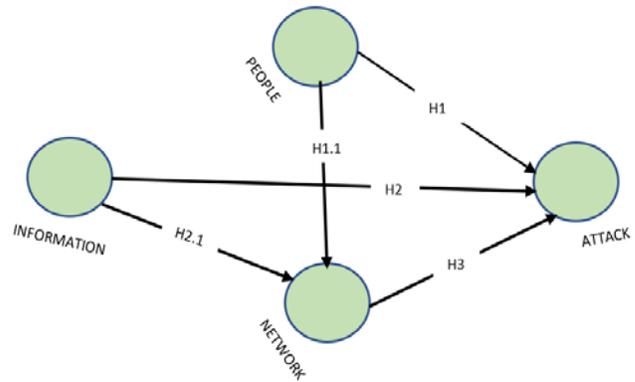


Fig. 4. Conceptual Framework.

Based on the above research hypotheses, the influence of cyber situational awareness on People, Information, and Network cyberattacks is present in Fig. 4.

## III. RESEARCH DESIGN AND METHODOLOGY

### A. Research Design

The approach of our current study is purely quantitative. We sort to use this approach because it involves an empirical investigation of the social phenomena of Cyber Attacks from Human Behavioral in the context of the IPPs. Therefore, the employs quantitative data using the survey method to collect and gather the information required for the analysis.

### B. Data Collection

Questionnaires were administered purposively to a sample of selected units of the IPPs staff who are directly involved in the organization's day-to-day grid operations from March to May 2021. We sort to use questionnaire because of anonymity among respondents and its low-cost implication. The questionnaire consists of four (4) parts: Demographic factor, People, Information, and Network. The questionnaire design is founded on [51] conceptual framework using a Likert scale from (One)1 for strongly disagree to (Five) 5 for strongly agree.

### C. Population and Sampling Procedures

The population studied is GRIDco and five (5) IPPs in Ghana that is actively operating with GRIDco in the power generation sector to assess Cyber Attacks on the electrical grid from human behaviour in context. Our estimated sample frame is 300, based on the various units within the IPPs: Supervisory Control and Data Acquisition (SCADA) Unit, Telecommunication Unit, and the Management Information Systems (MIS). The sampling method employed was probability sampling (random sampling). Choice of employees was at random within the various IPPs units selected to answer the questions

### D. Sample Size Calculation

Our Sample size is calculated based on Yamane 1967 formula with 95 per cent confidence level plus or minus 5 per cent confidence intervals using the formula  $n = \frac{N}{1+N(e)^2}$  where  $n$ = is the sample size  $N$  =is the population,

and  $e$  is the error margin. Although a sample size of 171 is obtained per our calculation from the estimated population of 300. Meanwhile, the Actual sample size used for the analysis is based on the number of respondents, which is (238) was used for the study because, with most research, a large sample size gives more reliable results than smaller samples.

#### IV. ANALYSIS AND RESULTS

The PLS-SEM approach seeks to support evaluating patterns of causality of target constructs in the structural model. It also provides more detailed statistical associations supporting variables included in a model. Authors' employees' PLS-SEM in the study since it allows relatively more minor samples than other statistical software like Amos and Lisrel. PLS-SEM does not enforce stringent assumptions on data distribution [52]. 238 is the valid responses received, which forms the basis of data analysis; 86.1% were males, while 13.9% were from females (see Table 2) on demographic of respondents. Using the two-step approach to evaluating structural equation models as suggested by [53], we began examining the measurement model to assess the instrument's reliability and validity. We then looked at the structural model based on the hypotheses proposed in this study.

##### A. Measurement Model

Following our measurement model assessment, we identified four(4) items (AK2, C5, C6, and C15) is removed from the study as a result of their low loadings, which is less than 0.600as recommended by ([54], [55]). In assessing the reliability of the constructs for our measurement Models, we used Cronbach's alpha and composite reliability measures to test the model's internal consistency. Hence Cronbach's alpha and composite reliability for each construct are adequate, as indicated in Table III. Composite reliability should be higher than 0.6 and 0.70 ([56]–[58]). With indicator factor loadings surpassing 0.5, Nunnally [58] and Hair, et al [59] recommended that the Average Variance Extracted for each variable should exceed 0.5 [56] to assess convergent validity [59]. Indications of convergent instrument validity are present

in Table III. Hence, convergent validity is suitable since the AVE values exceed 0.500 [60]. We strictly adhere to the Fornell–Larcker criterion in reporting the discriminant validity, which posits that AVE for each latent construct should be higher than the construct's highest squared correlation with any other latent construct [56].

Additionally, the loadings of each indicator should be greater than all its cross-loadings [53]. An inspection of indicator cross-loadings in Table III shows that all indicators load their highest on their respective construct. No Indicator loads higher on other constructs than on its intended construct. Evident in Table 4 shows that the square root of the AVEs for each construct is greater than the cross-correlation for different constructs. Based on these results is the established discriminant validity of the instrument. We also measured Discriminant validity using the Heterotrait-Monotrait Ratio of correlation(HTMT) criterion by [61], [62], see Table V below.

TABLE II. DEMOGRAPHY OF RESPONDENTS

| ATTRIBUTE           | CATEGORIES                 | PERCENTAGE (%) |
|---------------------|----------------------------|----------------|
| Gender              | Male                       | 86.1           |
|                     | Female                     | 13.9           |
| Age                 | 18 -30                     | 0.8            |
|                     | 31-40                      | 12.6           |
|                     | 41-50                      | 19.7           |
|                     | 51-60                      | 37.8           |
|                     | 61+                        | 29             |
| Education           | Diploma                    | -              |
|                     | HND                        | -              |
|                     | First degree               | 23.1           |
|                     | Postgraduate degree        | -              |
|                     | Professional qualification | 35.7           |
|                     | Masters                    | 41.7           |
| Years of experience | PhD                        | -              |
|                     | 1-5                        | 1.7            |
|                     | 6-10                       | 15.5           |
|                     | 11-15                      | 32.8           |
|                     | 16-20                      | 28.2           |
| 21 and above        | 21.8                       |                |

TABLE III. LOADINGS OF RELIABILITY AND VALIDITY RESULTS

|     | ATTACK       | INFORMATION  | NETWORK      | PEOPLE       | CA           | CR           | AVE          |
|-----|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| AK1 | <b>0.851</b> | 0.362        | 0.352        | -0.287       | <b>0.875</b> | <b>0.921</b> | <b>0.797</b> |
| AK3 | <b>0.896</b> | 0.233        | 0.215        | -0.509       |              |              |              |
| AK4 | <b>0.929</b> | 0.564        | 0.568        | -0.454       |              |              |              |
| C1  | -0.436       | -0.033       | -0.058       | <b>0.948</b> | <b>0.934</b> | <b>0.953</b> | <b>0.836</b> |
| C2  | -0.466       | -0.27        | -0.259       | <b>0.894</b> |              |              |              |
| C3  | -0.163       | 0.386        | 0.394        | <b>0.742</b> |              |              |              |
| C4  | -0.447       | -0.014       | -0.013       | <b>0.938</b> |              |              |              |
| C7  | 0.394        | <b>0.908</b> | 0.849        | -0.051       | <b>0.922</b> | <b>0.945</b> | <b>0.811</b> |
| C8  | 0.392        | <b>0.856</b> | 0.839        | -0.17        |              |              |              |
| C9  | 0.413        | <b>0.919</b> | 0.827        | -0.02        |              |              |              |
| C10 | 0.459        | <b>0.971</b> | 0.949        | -0.103       |              |              |              |
| C11 | 0.417        | 0.872        | <b>0.909</b> | -0.166       | <b>0.916</b> | <b>0.934</b> | <b>0.782</b> |
| C12 | 0.338        | 0.732        | <b>0.858</b> | -0.081       |              |              |              |
| C13 | 0.422        | 0.91         | <b>0.887</b> | -0.03        |              |              |              |
| C14 | 0.421        | 0.887        | <b>0.947</b> | -0.076       |              |              |              |

TABLE IV. DISCRIMINANT VALIDITY USING FORNELL–LARCKER CRITERION

|             | ATTACK | INFORMATION | NETWORK | PEOPLE |
|-------------|--------|-------------|---------|--------|
| ATTACK      | 0.892  |             |         |        |
| INFORMATION | 0.454  | 0.914       |         |        |
| NETWORK     | 0.446  | 0.949       | 0.901   |        |
| PEOPLE      | -0.475 | -0.094      | -0.097  | 0.885  |

Notes: Construct correlations with the square root of AVE along the diagonals

TABLE V. HETEROTRAIT-MONOTRAIT RATIO HTMT

|             | ATTACK | INFORMATION | NETWORK | PEOPLE |
|-------------|--------|-------------|---------|--------|
| ATTACK      |        |             |         |        |
| INFORMATION | 0.477  |             |         |        |
| NETWORK     | 0.469  | 1.017       |         |        |
| PEOPLE      | 0.47   | 0.241       | 0.247   |        |

**B. Structural Model**

In this section, overall explanatory power, Amount of variance explained by the independent variables, the degree of strength of each path is assessed. To estimate path significance, we applied bootstrap. We also assess the quality of the structural model with the coefficient of determination ( $R^2$ ) and standardized root mean square residual (SRMR) ([63] [64]). Our structural model results are present in Fig. 5 and Table 6. Regarding our (H1), The result shows even though people have a negative influence in predicting cyber-attack plays a significant role for an attack to occur ( $\beta = -0.435$ ;  $p = 0.000$ ). (H1.1) However, the mediation effect of network activities from people to attack does not have a negative relationship and plays no significant role in predicting cyber-attack within the grid. ( $\beta = -0.008$ ;  $p = 0.643$ ). (H2) our results reveal that information handling is significant and positively influences cyber-attack predictions within the grid. ( $\beta = 0.296$ ;  $p = 0.020$ ). Meanwhile, (H2.1) information handling positively affects network activities within the grid and significantly influences cyber-attacks ( $\beta = -0.949$ ;  $p = 0.000$ ). (H3) was found to have network activities positively affect predicting cyber-attack but not significant ( $\beta = 0.123$ ;  $p = 0.300$ ). Hence in this study, we did not find support for H3 and H1.1, as indicated in table 6 below. According to [62], the predictive validity of variance is a criterion for determining a model's prediction accuracy. Hence the coefficient of determination ( $R^2$ ) is the output of regression value as variance proportion in endogenous variable predicted by exogenous variable.

$R^2$  values range from 0 to 1; A higher value is said to have a higher level of  $R^2$  of .75 is substantial, .50 is moderate, and .25 is considered as weak ([65], [66]).

This study shows Attack (0.396, being Moderate) and Network with (0.901, substantial) value. In conclusion, the  $R^2$  indicates a sufficient level of  $R^2$  (see Table VI) and Fig. 5.

The authors performed analysis to assess the mediation role of the Network between People and Attack. The study results in (Table VII) reveal that the total effect of People on Attacks, even though negative is significant (H1.1:  $\beta = -0.436$ ,  $p = 0.000$ ). With the introduction of the mediator variable Network, the impact of People on Attack gives negative effect but significant ( $\beta = -0.435$ ,  $p = 0.000$ ). The indirect impact of people on Attacks through networks is insignificant ( $\beta = 0.116$ ,  $p = 0.318$ ), which indicates no mediations effect between the association of People and attacks.

Finally, assessing the mediation role of the Network between Information and Attack. Our results in (Table VII) also show that the total effect of Information on Attacks has a positive impact and is significant (H2.1:  $\beta = 0.413$ ,  $p = 0.000$ ). By introducing the mediator variable Network, the information effect on the Attack shows a positive impact and is significant ( $\beta = 0.296$ ,  $p = 0.022$ ). The indirect implications of Information on Attacks through networks have a positive effect and are insignificant ( $\beta = -0.001$ ,  $p = 0.768$ ), which indicates no mediations effect between the association of Information and Attack.

TABLE VI. PATH COEFFICIENTS AND THEIR SIGNIFICANCE

| Hypotheses | Path                   | Standardized path coefficient | T Statistics | P Values | Result        |
|------------|------------------------|-------------------------------|--------------|----------|---------------|
| H1         | PEOPLE -> ATTACK       | -0.435                        | 5.704        | 0.000    | Supported     |
| H1.1       | PEOPLE -> NETWORK      | -0.008                        | 0.463        | 0.643    | Not Supported |
| H2         | INFORMATION -> ATTACK  | 0.296                         | 2.330        | 0.020    | Supported     |
| H2.1       | INFORMATION -> NETWORK | 0.949                         | 91.394       | 0.000    | Supported     |
| H3         | NETWORK -> ATTACK      | 0.123                         | 1.037        | 0.300    | Not Supported |
|            |                        | <b>R<sup>2</sup></b>          |              |          |               |
|            | ATTACK                 | 0.396                         |              |          |               |
|            | NETWORK                | 0.901                         |              |          |               |

Notes: SRMR= 0.155; ns-not significant.

TABLE VII. MEDIATION RESULTS

|                       | TOTAL EFFECT     |           | DIRECT EFFECT    |           |                                 | INDIRECT EFFECT  |           |
|-----------------------|------------------|-----------|------------------|-----------|---------------------------------|------------------|-----------|
|                       | PATH COEFFICIENT | P-VALUE S | PATH COEFFICIENT | P-VALUE S |                                 | PATH COEFFICIENT | P-VALUE S |
| PEOPLE -> ATTACK      | -0.436           | 0.000     | -0.435           | 0.000     | PEOPLE -> NETWORK-> ATTACK      | 0.116            | 0.999     |
| INFORMATION -> ATTACK | 0.413            | 0.000     | 0.296            | 0.022     | INFORMATION -> NETWORK-> ATTACK | -0.001           | 0.296     |

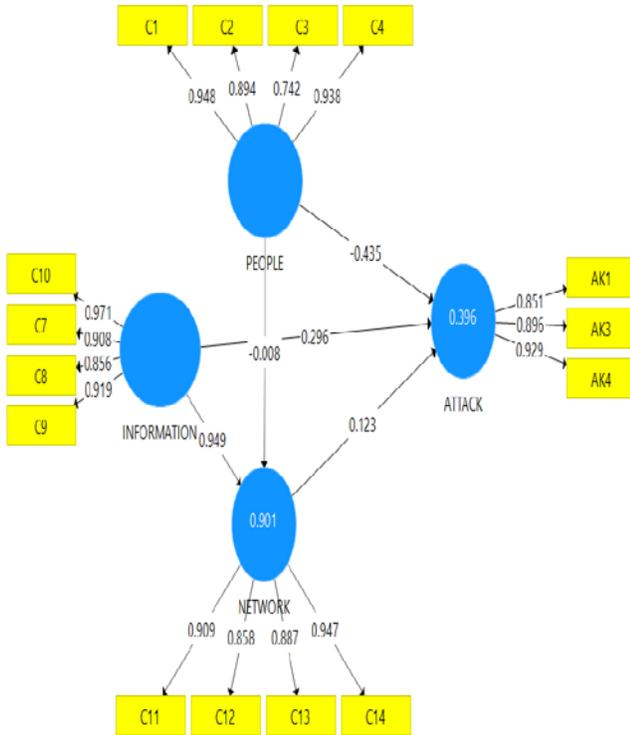


Fig. 5. PLS Results for Structural Model.

C. Effect Size ( $f^2$ )

Effect size is a concept to measure how strong the relationship of an indicator or the effect of exogenous constructs to endogenous constructs is. It examines changes in  $R^2$  value when an exogenous construct is detached from the model. An effect size of.02 has a minor influence, a value of.15 has a medium effect, and a value of.35 has a significant impact. [66]. Hence the current study revealed four correlational effect sizes. C15 with a value of 0.018 as a small effect, AK2 and C6 also gave indications of values of 0.264 and 0.543, respectively, having an effect size of the medium impact.C5 recorded the most significant effect sizes with the value of 0.617.

V. IMPLICATIONS OF THE STUDY

CSA aims to improve the quality of appropriateness of concerted decision-making concerning the protection of the electrical grid. Invalidating our hypothesis, although (H1) posited there is a negative effect of people in predicting cyber-attack, it also plays a significant role for an attack to occur in the operational environment of the IPPs. Therefore, consider

cyber security measures to ensure operational staff knows potential external cyber threats. Ensuring that the grid and its sub-systems are fully protected, prohibiting personal devices from accessing and transmitting information on the grid. Finally, steps must be taken to monitor users' grid activities strictly. Hypothesis (H1.1) indicates no mediation effect between the association of People and attacks. Explaining that, irrespective of users' activities such as personal devices, remote access to the network by I.T and Non-I.T staff can prevent an attack on the electrical grid network with the proper security controls. Frequent changes in network access security policies (H2) suggest that information handling is significant and positively influences cyber-attack. Therefore, lack of CSA from the perspective of information accessibility and distribution concerning cyber security controls will distort confidential and sensitive information; hence the need to ensure the information on the grid activities is well-coordinated and accessible to only specific users within the operational environment. The mediation effect of H2.1 indicates no mediations effect between the association of Information and Attack. Reasons could be due to an effective cyber security control is implemented by the IPPs. Information from the grid received with inaccuracies is attributed to frequent updates of security policies in the context of perspective network access. (H3) reveals no sign of cyber-attack predictions. It indicates that the grid infrastructure network managers have put adequate security measures. Such as frequent network access policies, hence irrespective of who performs activities on the network do not translate into any negative impact on the grid network.

VI. CONCLUSION

Over the years, the power generation and distribution sectors have seen an increase in the number of independent power producers in the electricity market. This sector of power generation has advanced its operations with the intervention of internet of things (IoT) technology and other electronic equipment making the entire grid network susceptible to attack [1], [67], [68]. From the literature perspective, very little is known about user behaviour concerning cyber vulnerabilities in the sector. Bradley [69] insider threats are the most expensive threats challenging to address threats to people, information, and technology in the business environment. In this regard, our research helps to improve the understanding of user behaviour in the context of Ghana's electricity generation sector. From the context of cyber situational awareness (CSA), authors sort to (1) evaluate how the operational staff of IPPs perceive cyber vulnerabilities within their operational environment, (2) evaluate the information vulnerabilities that seeks to influence cyber-attack and finally (3) assess network

vulnerabilities that contribute to cyber-attack on the electrical grid. Our findings show that People's construct negatively predicts cyber-attack but plays a significant role in attacking the electrical grid. The authors also realized no mediation effect from people and attack network activities, probably because of constant security controls measures such as frequent update network access policies. Meanwhile, authors also realized that information handling is significant and positively influences cyber-attack, which calls for well-coordinated cyber security controls on the grid activities in line with confidential and sensitive information in the operational environment of IPPs. In addition, there was no indication of mediations effect from network activities between the association of Information and Attack. Such development can be to adequate security measures being put in place to ensure information from the grid are received devoid of errors. Finally, managers of IPPs infrastructure networks seem to have suitable security measures concerning the network activities. Hence irrespective of the numerous activities on the network cannot easily translate into any negative cyber-attack impact on the grid network.

#### REFERENCES

- [1] T. R. Sharafiev, O. V Ju, and A. L. Kulikov, "Cyber-Security Problems in Smart Grid," 2018 Int. Conf. Ind. Eng. Appl. Manuf., pp. 1–6, 2018.
- [2] M. A. Shahid, R. Nawaz, I. M. Qureshi, and M. H. Mahmood, "Proposed Defense Topology against Cyber Attacks in Smart Grid," 4th Int. Conf. Power Gener. Syst. Renew. Energy Technol. PGSRET 2018, no. September, pp. 1–5, 2019, DOI: 10.1109/PGSRET.2018.8685944.
- [3] R. J. Campbell, "Electric Grid Cybersecurity," Congr. Res. Serv., 2018, [Online]. Available: <https://crsreports.congress.gov>.
- [4] H. Jia, C. Shao, S. Member, and D. Liu, "Operating Reliability Evaluation of Power Systems With Demand-Side Resources Considering Cyber Malfunctions," IEEE Access, vol. 8, 2020, DOI: 10.1109/ACCESS.2020.2992636.
- [5] T. Nguyen and S. Wang, "Electric Power Grid Resilience to Cyber Adversaries: State of the Art," IEEE Access, vol. 8, 2020, DOI: 10.1109/ACCESS.2020.2993233.
- [6] V. S. Rajkumar, M. Tealane, and S. Alexandru, "Cyber Attacks on Protective Relays in Digital Substations and Impact Analysis," IEEE Xplore, 2020.
- [7] U. Franke and J. Brynielsson, "Cyber situational awareness - A systematic review of the literature," Comput. Secure., vol. 46, pp. 18–31, 2014, DOI: 10.1016/j.cose.2014.06.008.
- [8] T. Nguyen, S. Wang, M. Alhazmi, M. Nazemi, A. Estebarsari, and P. Dehghanian, "Electric Power Grid Resilience to Cyber Adversaries: State of the Art," IEEE Access, vol. 8, pp. 87592–87608, 2020, DOI: 10.1109/ACCESS.2020.2993233.
- [9] Z. Zhang, "Cybersecurity Policy for the Electricity Sector: The First Step to Protecting our Critical Infrastructure from Cyber Threats," J. Sci. Technol. Law, vol. 19, no. 2, pp. 319–366, 2013.
- [10] H. He, S. Huang, Y. Liu, and T. Zhang, "International Journal of Electrical Power and Energy Systems A tri-level optimization model for power grid defense with the consideration of post-allocated DGs against coordinated cyber-physical attacks," Int. J. Electr. Power Energy Syst., vol. 130, no. March, p. 106903, 2021, DOI: 10.1016/j.ijepes.2021.106903.
- [11] P. Johnson and D. Z. Banfield, "Energy Security Forum," Q. J., vol. 3, no. 6, pp. 1–12, 2012.
- [12] S. Jajodia, P. Liu, V. Swarup, and C. Wang, Cyber situational awareness: advances in information security. 2010.
- [13] M. N. Lakhous, "Cyber Security of SCADA Network in Thermal Power Plants," 2018 Int. Conf. Smart Commun. Netw., pp. 1–4, 2018.
- [14] R. K. Pandey, "Cyber Security Threats - Smart Grid Infrastructure," 2016.
- [15] A. Janjić, L. Velimirović, J. Ranitović, and Ž. Džunić, "Internet of Things in Power Distribution Networks – State of the Art," no. September 2017.
- [16] A. Ramamurthy and P. Jain, "The Internet of Things in the Power Sector Opportunities in Asia and the Pacific," no. 48, 2017.
- [17] C. A. Jones, G. Runger, and J. Caravelli, "HUMAN BEHAVIOUR AND DIGITAL TRUST HUMAN BEHAVIOUR AND DIGITAL TRUST :," pp. 1–6, 2017.
- [18] M. R. Endsley, "Human Factors : The Journal of the Human Factors and Ergonomics Society," 1995, DOI: 10.1518/001872095779049543.
- [19] U. States and G. Accountability, "CRITICAL INFRASTRUCTURE Actions Needed to Address Significant Cybersecurity Risks Facing the Electric Grid," no. August 2019.
- [20] J. C. Creasey, "Protecting critical national infrastructure through collaborative cyber situational awareness," IET Conf. Publ., vol. 2013, no. 620 CP, pp. 1–4, 2013, DOI: 10.1049/cp.2013.1708.
- [21] Flick T. and Morehouse J., "Securing the Smart Grid: Next Generation Power Grid Security," Syngress, 2010.
- [22] Y. Zhang, "Model-Based and Data-driven Situational Awareness for Distribution System Monitoring and Control," 2020.
- [23] V. A. Nekha and M. Dorosh, "Using the Cyber Situational Awareness Concept for Protection of Agricultural Enterprise Management Information Systems," vol. 46, no. 2, pp. 168–181, 2020.
- [24] T. Jirsik and P. Celeda, "Cyber Situation Awareness via IP Flow Monitoring," 2020.
- [25] X. Wei, C. Du, and J. Zhao, "A network security situation awareness model for electric vehicle shared charging pile system A Network Security Situation Awareness Model for Electric Vehicle Shared Charging Pile System," vol. 020009, no. May 2020.
- [26] A. H. Matey, P. Danquah, and G. Y. K.-A. I. Asampana, "CRITICAL INFRASTRUCTURE CYBERSECURITY CHALLENGES: IOT IN PERSPECTIVE," vol. 13, no. 4, 2021, DOI: 10.5121/ijnsa.2021.13404.
- [27] S. D. Roy and S. Debbarma, "Detection and Mitigation of Cyber-Attacks on AGC Systems of Low Inertia Power Grid," IEEE Syst. J., vol. 14, no. 2, pp. 2023–2031, 2020, doi: 10.1109/JSYST.2019.2943921.
- [28] P. Eder-Neuhauser, T. Zseby, J. Fabini, and G. Vormayr, "Sustainable Energy, Grids and Networks Cyberattack models for smart grid environments," Sustain. Energy, Grids Networks, vol. 12, pp. 10–29, 2017, DOI: 10.1016/j.segan.2017.08.002.
- [29] S. Jajodia, P. Liu, V. Swarup, and C. Wang, "Cyber Situational Awareness: Issues and Research," Springer, 2010.
- [30] M. Eckhart, A. Ekelhart, and E. Weippl, "Enhancing Cyber Situational Awareness for Cyber-Physical Systems through Digital Twins," 2019 24th IEEE Int. Conf. Emerg. Technol. Fact. Autom., vol. 1, pp. 1222–1225, 2019.
- [31] S. Pournouri, "Improving Cyber Situational Awareness via Data mining and Predictive Analytic Techniques," 2019.
- [32] S. J. Michael A. Champion, Prashanth Rajivan, Nancy J. Cooke, "Team-Based Cyber Defense Analysis. In 2012 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)," 2012.
- [33] O. Jacq, D. Brosset, Y. Kermarrec, and J. Simonin, "Cyber attacks real-time detection: towards a Cyber Situational Awareness for naval systems," 2019 Int. Conf. Cyber Situational Awareness, Data Anal. Assess. (Cyber SA), pp. 1–2, 2019.
- [34] V. V. G. Krishnan, Y. Zhang, K. Kaur, A. Hahn, A. Srivastava, and S. Sindhu, "Cyber-Security Analysis of Transactive Energy Systems," 2018 IEEE/PES Transm. Distrib. Conf. Expo., pp. 1–9, 2018.
- [35] Y. Wang, "analysis of electric cyber-physical systems," 2015.
- [36] M. Sahabuddin, B. Dutta, and M. Hassan, "Impact of cyber-attack on the isolated power system," 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. iCEEICT 2016, pp. 8–11, 2017, DOI: 10.1109/CEEICT.2016.7873088.
- [37] U. Javed Butt, M. Abbod, A. Lors, H. Jahankhani, A. Jamal, and A. Kumar, "Ransomware Threat and its Impact on SCADA," Proc. 12th Int. Conf. Glob. Secure. Saf. Sustain. ICGS3 2019, pp. 205–212, 2019, DOI: 10.1109/ICGS3.2019.8688327.

- [38] M. Ashrafuzzaman, S. Das, Y. Chakhchoukh, S. Shiva, and F. T. Sheldon, "Computers & Security Detecting stealthy false data injection attacks in the smart grid using ensemble-based machine learning," *Comput. Secure.*, vol. 97, p. 101994, 2020, DOI: 10.1016/j.cose.2020.101994.
- [39] X. Li and K. W. Hedman, "Enhancing Power System Cyber-Security with Systematic Two-Stage Detection Strategy," *IEEE Trans. Power Syst.*, vol. 35, no. 2, pp. 1549–1561, 2020, DOI: 10.1109/TPWRS.2019.2942333.
- [40] A. Sadu, A. Jindal, G. Lipari, F. Ponci, and A. Monti, "Resilient Design of Distribution Grid Automation System against cyber-physical attacks using Blockchain and Smart Contract," *Blockchain Res. Appl.*, p. 100010, 2021, DOI: 10.1016/j.bcra.2021.100010.
- [41] S. N. Narayanan, K. Khanna, and B. K. Panigrahi, *Security in Smart Cyber-Physical Systems : A Case Study on Smart Grids and Smart Cars*. Elsevier Inc., 2019.
- [42] M. Snehi and A. Bhandari, "Vulnerability retrospection of security solutions for software-defined Cyber-Physical System against DDoS and IoT-DDoS attacks," *Comput. Sci. Rev.*, vol. 40, p. 100371, 2021, DOI: 10.1016/j.cosrev.2021.100371.
- [43] L. Arnaboldi, R. M. Czekster, C. Morisset, and R. Metere, "Modelling Load-Changing Attacks in Cyber-Physical Systems," *Electron. Notes Theor. Comput. Sci.*, vol. 353, pp. 39–60, 2020, DOI: 10.1016/j.entcs.2020.09.018.
- [44] P. Matoušek, O. Ryšavý, M. Grégr, and V. ech Havlena, "Journal of Information Security and Applications Flow-based monitoring of ICS communication in the smart grid," *J. Inf. Secure. Appl.*, vol. 54, 2020, DOI: 10.1016/j.jisa.2020.102535.
- [45] T. Zou, A. S. Bretas, C. Ruben, S. C. Dhulipala, and N. Bretas, "Smart grids cyber-physical security: Parameter correction model against unbalanced false data injection attacks ☆," *Electr. Power Syst. Res.*, vol. 187, no. June, p. 106490, 2020, DOI: 10.1016/j.epr.2020.106490.
- [46] Franke U and J. Brynielsson, "Cyber situational awareness a systematic review of the literature," *Comput. Secure.*, vol. 46, pp. 18–31, 2014.
- [47] A. Zimba, Z. Wang, and H. Chen, "Multi-stage crypto-ransomware attacks: A new emerging cyber threat to critical infrastructure and industrial control systems," *ICT Express*, vol. 4, no. 1, pp. 14–18, 2018, DOI: 10.1016/j.icte.2017.12.007.
- [48] S. N. Islam, M. A. Mahmud, and A. M. T. Oo, "Impact of optimal false data injection attacks on local energy trading in a residential microgrid," vol. 4, pp. 30–34, 2018, DOI: 10.1016/j.icte.2018.01.015.
- [49] T. Akhtar, B. B. Gupta, and S. Yamaguchi, "Malware propagation effects on SCADA system and Smart Power Grid," 2018.
- [50] S. Sarangan, V. K. Singh, and M. Govindarasu, "Cyber Attack-Defense Analysis for Automatic Generation Control with Renewable Energy Sources," 2018 North Am. Power Symp., no. December 2015, pp. 1–6, 2018.
- [51] Fairclough and Happa, *Cyber Warfare and Organised Crime. A Regulatory Model and Meta-Model for Open Source Intelligence (OSINT)*. 2017.
- [52] W. C. Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, "Multivariate data analysis," Up. Saddle River, NJ Prentice-Hall, 1998.
- [53] W. Chin, "The Partial Least Squares Approach to Structural Equation Modeling. Modern Methods for Business Research," no. 295, p. 33, 1998.
- [54] J. F. Hair, G. T. M. Hult, and C. M. Ringle, *A Primer on Partial Least Squares Structural Equation Modeling ( PLS-SEM )*. 2017.
- [55] D. Gefen and D. Straub, "A Practical Guide To Factorial Validity Using PLS- Graph : Tutorial And Annotated Example," vol. 16, no. July 2005, DOI: 10.17705/1CAIS.01605.
- [56] D. Fornell, C. & Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *J. Mark. Res.*, vol. 18, no. 1, pp. 39–50, 1981.
- [57] J. Hair, C. L. Hollingsworth, A. B. Randolph, and A. Y. L. Chong, "An updated and expanded assessment of PLS-SEM in information systems research," *Ind. Manag. Data Syst.*, vol. 117, no. 3, pp. 442–458, 2017, DOI: 10.1108/IMDS-04-2016-0130.
- [58] J. C. Nunnally, "Psychometric theory," New York, NY McGraw-Hill., 1978.
- [59] R. L. Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, "Multivariate data analysis," New Jersey, NJ Prentice-Hall, 2006.
- [60] J. F. Hair, C. M. Ringle, and M. Sarstedt, "PLS-SEM: Indeed a silver bullet," *J. Mark. Theory Pract.*, vol. 19, no. 2, pp. 139–152, 2011, DOI: 10.2753/MTP1069-6679190202.
- [61] J. Henseler, G. Hubona, and P. A. Ray, "Using PLS path modelling in new technology research : updated guidelines," vol. 116, no. 1, pp. 2–20, 2016, DOI: 10.1108/IMDS-09-2015-0382.
- [62] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modelling," *J. Acad. Mark. Sci.*, vol. 43, no. 1, pp. 115–135, 2015, DOI: 10.1007/s11747-014-0403-8.
- [63] J. F. Hair, J. J. Risher, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," vol. 31, no. 1, pp. 2–24, 2018, DOI: 10.1108/EBR-11-2018-0203.
- [64] R. J. Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A. Straub, D. W., Calantone, "Common beliefs and reality about PLS: Comments on Ronkko and Evermann (2013). *Organizational Research Methods*," vol. 182–209, no. 17, p. 2, 2014.
- [65] G. Shmueli et al., "Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict," *Eur. J. Mark.*, vol. 53, no. 11, pp. 2322–2347, 2019, DOI: 10.1108/EJM-02-2019-0189.
- [66] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," *Eur. Bus. Rev.*, vol. 31, no. 1, pp. 2–24, 2019, DOI: 10.1108/EBR-11-2018-0203.
- [67] A. Ghasempour, "Internet of Things in Smart Grid: Architecture, Applications, Services, Key Technologies, and Challenges," 2019, doi: 10.3390/inventions4010022.
- [68] H. P. Tauqir and A. Habib, "Integration of IoT and Smart Grid to Reduce Line Losses," 2019 2nd Int. Conf. Comput. Math. Eng. Technol., pp. 1–5, 2019.
- [69] P. Bradley, "THE INSIDER SECURITY THREAT," *cyber Secure. Rev.*, 2016.

APPENDIX 1

Instrument: Cyber Situational Awareness(Csa)

SECTION A: DEMOGRAPHIC (Please tick appropriately)

1) Please indicate your gender:

Male [ ]

Female [ ]

2) Please indicate your age category:

18 -30 [ ]

31-40 [ ]

41-50 [ ]

51-60 [ ]

61 [ ]

3) Please indicate your highest category:

- Diploma
- HND
- First degree
- Postgraduate degree
- Professional qualification
- Masters
- PhD

4) Please indicate your years of experience on the job:

- 1-5
- 6-10
- 11-15
- 16-20
- 21 and above

SECTION B: Please tick appropriately

| Please tick the correct numeric response to each question |                                                                                                      | 1= Strongly Disagree, 2=Disagree, 3=Neutral, 4= Agree, 5= Strongly Agree |   |   |   |   |
|-----------------------------------------------------------|------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|---|---|---|---|
|                                                           | <b>People</b>                                                                                        | 1                                                                        | 2 | 3 | 4 | 5 |
| C1                                                        | I am aware of potential cyber threats from external sources on our network                           |                                                                          |   |   |   |   |
| C2                                                        | I am aware that our systems are well secured                                                         |                                                                          |   |   |   |   |
| C3                                                        | I know my colleagues use their devices on our network                                                |                                                                          |   |   |   |   |
| C4                                                        | It is difficult to control other computer users on the network.                                      |                                                                          |   |   |   |   |
| C5                                                        | my colleagues and I always strictly follow our cyber security protocols                              |                                                                          |   |   |   |   |
|                                                           | <b>Information</b>                                                                                   |                                                                          |   |   |   |   |
| C6                                                        | We do not apply cyber security controls when interacting with confidential and sensitive information |                                                                          |   |   |   |   |
| C7                                                        | We usually experience leakages of vital and sensitive information                                    |                                                                          |   |   |   |   |
| C8                                                        | I have access to data remotely without restriction                                                   |                                                                          |   |   |   |   |
| C9                                                        | The Information sometimes receive from the grid comes with inaccuracies due to cyber Attack          |                                                                          |   |   |   |   |
| C10                                                       | We frequently receive misleading information from our systems                                        |                                                                          |   |   |   |   |
|                                                           | <b>Network</b>                                                                                       | 1                                                                        | 2 | 3 | 4 | 5 |
| C11                                                       | I can access the network with my devices                                                             |                                                                          |   |   |   |   |
| C12                                                       | I am aware staff can access social media applications on the network                                 |                                                                          |   |   |   |   |
| C13                                                       | I am aware unauthorized non-IT staff can access the network remotely.                                |                                                                          |   |   |   |   |
| C14                                                       | I am aware unauthorized IT staff can access the network remotely.                                    |                                                                          |   |   |   |   |
| C15                                                       | Network access policies changes frequently                                                           |                                                                          |   |   |   |   |
|                                                           | <b>Attack</b>                                                                                        | 1                                                                        | 2 | 3 | 4 | 5 |
| AK1                                                       | There is the manipulation of other components on the grid by attackers                               |                                                                          |   |   |   |   |
| AK2                                                       | Attackers take advantage of cyber vulnerabilities in the IoT devices                                 |                                                                          |   |   |   |   |
| AK3                                                       | We experience loss of generations capability disabling power generation network by attackers         |                                                                          |   |   |   |   |
| AK4                                                       | I have access to all systems resource                                                                |                                                                          |   |   |   |   |

# Design and Performance Analysis of Anti-Surge Control Mechanism for Compressor System using Neural Networks

Divya M.N<sup>1</sup>

Research Scholar  
VTU Research Centre, MSRIT  
VTU, Belagavi, India

S L Gangadhariah<sup>3</sup>

Dept. of Electronics and Communication  
M.S. Ramaiah Institute of Technology  
Bangalore, India

Narayanappa C.K<sup>2</sup>

Dept. of Medical Electronics  
M.S. Ramaiah Institute of Technology, Bangalore, India

V Nuthan Prasad<sup>4</sup>

Dept. of Electronics and Communication Organization  
M.S. Ramaiah Institute of Technology, Bangalore, India

**Abstract**—The compressor system is caused by the surge, which is an instability occurrence in most gas-process and oil industries. These issues are solved by using a recycle valve that avoids the surge and provides higher mass flow in the compressor system. An advanced controller-based anti-surge control mechanism is a need in the compressor system to improve the stability and surge issues. In this manuscript, an efficient, Neural-network predictive controller (NNPC) based variable speed compressor recycle system is modeled with an anti-surge control mechanism. When the mass flow is deficient, the recycle system is introduced, acts as a safety system, and feeds the compressed gas back to the upstream system. The different controllers like Proportional Integral Derivative (PID) controller, Fuzzy logic controller (FLC), and Neuro-fuzzy controller (NFC) based anti-surge control mechanism are also used in Compressor recycle system to compare the stability and performance metrics with NNPC. The NNPC based compressor system provides a better operating position and dynamic response with less error than other controllers-based compressor systems.

**Keywords**—Anti-surge; fuzzy logic controller (FLC); neuro-fuzzy controller (NFC); compressors; neural-network

## I. INTRODUCTION

The centrifugal Compressor is used in most industries like oil, aero-space, and gas-plant to increase pressure and oil production. The gas compressors are generally divided into four types: Axial-flow, centrifugal, rotary, and Reciprocating-type. The rotary and Reciprocating-type compressors reduce the gas volume's occupancy and later apply the higher Pressure to discharge the gas. The axial and centrifugal compressors act as turbo-compressor with continuous flow [1-2]. The Axial compressors with rotating stall and surge are presented with theoretical and mathematical concepts by Greitzer et al. [3]. The non-dimensional parameter (B) factors define whether the Compressor is in surge or rotating stall conditions. If the 'B' factor is lower, then a rotating stall will occur, which causes the instabilities by reducing the mass flow by rising the Pressure.

Similarly, a surge will occur with the mass flow and pressure oscillations if the 'B' factor is higher. The spool dynamics are introduced in the 'B' parameter to improve the speed control [4]. The Full-range Surge and rotating-stall avoidance by Badmus et al. are presented in the compressor system. The model uses open-loop feed-forward and closed-loop feedback control mechanisms in the throttle valve area to control the Pressure and mass flow. These control mechanisms are scheduled and use in rotating stall and surge avoidance [5]. The calculation of operating location between surge and the stable condition is necessary to avoid the surge in Compressor. To counteract the surge, the invariant coordinate system is introduced by changing molecular weight [6]. Gravdahl et al. present the drive-torque actuation system to regulate the active surge in the centrifugal Compressor. The active surge stabilization utilizing control law [7] causes the throttle line to appear left to the surge line.

In contrast, the compressor system for active surge control is presented by bohagen et al. using drive torque as a control input. The desired operating point stabilization is achieved with static and dynamic feedback mechanisms [8]. The Fuzzy logic-based surge control mechanism [9] is introduced for the Moore-Greitzer model with constant speed in the compressor system, stabilizing the different operating conditions and extending the stable line next to the surge line. The active surge control with variable speed mechanism [10] in centrifugal Compressor is designed to analyze the different performance metrics with better improvement than the Gravdahl et al. [1] work. The Contribution of the research work is organized as follows:

- The complete variable speed-based Compressor system with a recycling loop is modeled to ensure higher mass flow and avoid the surge.
- The Neural-network predictive controller (NNPC) based anti-surge control mechanism is modeled for the Compressor recycle system to improve stability and prevent active surges.

- The different controllers like PID controller, Fuzzy logic controller (FLC), and Neuro-fuzzy controller (NFC) are also modeled in anti-surge control mechanism to analyze the stability and performance comparison.

In this manuscript, an efficient Compressor recycle system with an anti-surge control mechanism using NNPC is modeled and compared with other controllers to evaluate performance metrics. The manuscript is organized as follows: Section II explains the Compressor-based system's current works for different applications. The simple compressor system with mathematical expressions is described in Section III. The Compressor recycle system with an anti-surge control mechanism using other controllers is explained in Section IV. The results and discussion is carried out in Section V. Section VI concludes the overall work with performance improvement and also suggest the future scope.

## II. RELATED WORK

This section analyzes the recent works on Compressor based systems for different application requirements. Budinis et al. [11-12] describe the Centrifugal compressor's control mechanism using a Model predictive controller (MPC), which can control the Compressor's Pressure during surge operation. The hot and cold recycle line is configured with an anti-surge control mechanism to analyze the performance metrics and power consumption. Cortinovis et al. [13-14] present the Linearize MPC-based centrifugal gas compressor system with an anti-surge and process control mechanism. These mechanisms are adopted in safe and electric-driven applications with improved surge 11% by maintaining distance and 50% reduction in process control time (settling). Sheng et al. [15] describe the Throttle and closed-coupled Valve (CCV) actuators based Compressor system for the instability control mechanism. The double actuators scheme is introduced to improve the control performance in CCV by tuning the B-parameter in the Throttle valve. Gritli et al. [16] present the Electronic throttle valve (ETV) optimization using a genetic algorithm (GA) by tuning the PID-based Fuzzy factors. The GA with PID-based FLC provides better disturbance rejection and trajectory tracking points than other conventional ETV approaches. Azeem et al. [17] present the ETV with FLC based super-twisting-Sliding mode control (STSMC) scheme. The STSMC scheme improves the controller's dynamic performance in ETV without compromising the tracking accuracy and stability features. Saeed et al. [18] present the Hybrid electric vehicles (HEVs) speed control mechanism using different controllers. The state observer controller (OBC) and linear quadratic regulator (LQR) based scheme provide better dynamic control response than the conventional PID Controller scheme in HEV.

The Neuro-Fuzzy controller (NFC) based Grid-connected gas turbine is modeled by Iqbal et al. [19] to improve the transient response and maintain stable operations. Liying et al. [20] present the non-linear compensation control mechanism using a PID controller for ETV to analyze the response time and performance. Zaibari et al. [21] explain the constant-speed-based centrifugal Compressor using Tube-MPC to control the surge instability. The Adaptive fuzzy logic with MPC acts as

an anti-surge controller to avoid the surge and improve the robustness against the mass flow and pressure-based disturbances. Khsheem et al. [22] present the Surge compressor system with an active control mechanism. The PID controller controls the surge line and provides the dynamic response to use in the control valve. Guan et al. [23] present the centrifugal compressor surge control mechanism. The work analyzes the variable –tip clearance mechanism is introduced to control the surge and analyzes the compressor performance. The Surge control with different speeds and also with different throttle opening sizes are analyzed with simulation results. Aribi et al. [24] present the Recycle compressor system for active surge control using a hybrid adaptive controller. The PID with auto tuning Fuzzy interface system (FIS) controls the operating points under the Surge line at different speeds. The artificial neural network (ANN) based Centrifugal compressor is modeled by Ebrahimi et al. [25] to predict operational parameters.

## III. COMPRESSOR MODEL

The surge control and instabilities are the main issues in any of the axial and centrifugal compressors. In this work, variable-speed Centrifugal-based Compressor is discussed. The overview of the Compressor system is represented in Fig. 1. It mainly contains a Torque drive, Compressor with duct, Plenum, and Throttle valve. The Compressor model produces the mass and Pressure using torque drive and plenum units. The Throttle valve delivers the mass throttle value and feedback to the Plenum unit and Compressor mass flow.

When the mass is applied to the Plenum unit, which produces the Pressure, it is represented in Equation (1). The momentum applied on the Compressor with duct produces the mass flow is described in Equation (2). The torque drive delivers the angular speed using the angular momentum feature represented in Equation (3). The work was initially presented by Gravdahl et al. [1], and it is extended by Fink et al. [2]. The Pressure, mass flow, and speed is represented in Equation (1-3) as follows:

$$\dot{P}_p = \frac{a_p^2}{V_p} (m - m_t) \quad (1)$$

$$\dot{m} = \frac{A}{L} (\psi_c(m, \omega) P_{01} - P_p) \quad (2)$$

$$\dot{\omega} = \frac{1}{J} (\tau_d - \tau_c) \quad (3)$$

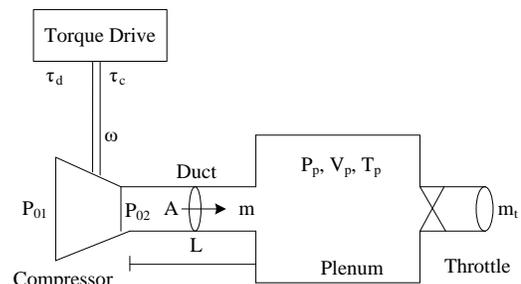


Fig. 1. An Overview of the Compressor System.

The above Equations (1-3) uses few of the notations as follows: where  $\dot{m}$  is mass flow,  $\dot{P}_p$  is Plenum Pressure,  $\omega$  is Torque speed of the impeller,  $V_p$  is Plenum Volume,  $a_p$  Plenum's speed of the gas,  $m_t$  is Throttle's mass flow,  $A$  is the cross-sectional area of the duct,  $L$  is the length of the duct,  $P_{01}$  is the Pressure (Ambient),  $\psi_c$  is Compressor Characteristic,  $J$  represents the moment of inertia of torque,  $\tau_d$  is Torque drive, and  $\tau_c$  is impeller blades torque.

The predicting surge is possible using Equations (1-3) and is one-dimensional in centrifugal compressors. When Compressor is in the deep surge point, and the impeller blade's torque is represented in Equation (4) as follows:

$$\tau_c = |m| r_2^2 \mu \omega \quad (4)$$

The Compressor system is used to determine the pressure ratio using  $\psi_c$ . The compressor characteristic is expressed in Equation (5) as follows [1]:

$$\psi_c(m, \omega) = \left( 1 + \frac{\mu r_2^2 \omega^2 - \frac{1}{2} r_1^2 (\omega - \alpha m)^2 - k_f m^2}{a_p T_1} \right)^{\frac{k}{k-1}} \quad (5)$$

The throttle mass flow is represented in Equation (6) as follows:

$$m_t = \tanh(\zeta(P_p - P_{01})) A_t \sqrt{(P_p - P_{01}) \tanh(\zeta(P_p - P_{01}))} \quad (6)$$

The above Equations (4-6) uses a few of the notations as follows:  $r_1$  is inducer radius, and  $r_2$  is impeller exit radius,  $\mu$  is flow co-efficient,  $k$  is heat capacity ratio,  $T_1$  is inlet temperature,  $A_t$  is Orifice Opening area, and  $\zeta$  is zeta ( $\gg 1$ ). The Compressor characteristic provides the performance metrics of the Compressor using mass flow and Pressure. When the mass flow decreases, the Pressure will increase until its instability point, Then Surge or rotating stall problems arise in the compressor system.

#### IV. COMPRESSOR WITH ANTI-SURGE CONTROL USING DIFFERENT CONTROLLERS

The oscillations, stalls/surge, and lower pressure rises will always affect the centrifugal Compressor's performance. The operating point is always far away from the surge line (SL), which is the best option at low pressure and high mass flow conditions. But the efficiency of the Compressor always lacks in this region. So for better efficiency, the Pressure will rise at its highest, which is close to the SL. Surge avoidance is also known as an anti-surge control mechanism. Most commonly, the recycle valves are used to avoid the surge in industry applications. The Compressor Recycle system based Anti-surge Control Mechanism using different Controllers is represented in Fig. 2.

Using an anti-surge control mechanism, the Compressor recycle system contains a Suction unit, Compressor with duct, Torque drive unit, Plenum and recycle valve units. The feed flow mechanism is essential to analyze the compressor system performance if the feed flow value is high or too low, which

affects the operating point and speed features in the compressor system. Suppose the pressure  $P_1$  value is too high in the suction unit, which fails to operate in the compressor system. The Suction unit with Pressure ( $P_1$ ), and Volume1 ( $V_1$ ), are part of the upstream piping operation. Similarly, The Plenum unit with Pressure ( $P_2$ ), and Volume2 ( $V_2$ ), are part of the downstream piping operation. The modeling of the Compressor recycle system is represented in Fig. 3.

The amount of feed flow ( $m_f$ ) with mass applied on the Suction unit, which generates the Pressure ( $P_1$ ) and is represented in Equation (7). The mass flow is used on the Plenum unit, which produces the Pressure ( $P_2$ ) and is represented in Equation (8). The momentum applied to the Compressor creates the mass flow ( $m$ ), described in Equation (9). The throttle mass flow of the compressor recycle system is represented in Equation (10) as follows:

$$\dot{P}_1 = \frac{a_p^2}{V_1} (m_f - m_r - m) \quad (7)$$

$$\dot{P}_2 = \frac{a_p^2}{V_2} (m - m_t - m_r) \quad (8)$$

$$\dot{m} = \frac{A}{L} (\psi_c(m, \omega) P_1 - P_2) \quad (9)$$

$$m_t = \tanh(\zeta(P_2 - P_{01})) C_t \sqrt{(P_2 - P_{01}) \tanh(\zeta(P_2 - P_{01}))} \Big|_{\zeta \gg 0} \quad (10)$$

The feed flow mechanism is dependent more on the type of Compressor used and is not constant while performing the simulation process. The feed flow ( $m_f$ ) operation is represented in Equation (11), and recycle flow ( $m_r$ ) operation is described in Equation (12) and are as follows:

$$m_f = C_f \sqrt{P_{01u} - P_1} \quad (11)$$

$$m_r = C_r \sqrt{P_2 - P_1} \quad (12)$$

$C_f$  and  $C_r$  are coefficients of Feed flow and recycle flow, respectively;  $P_{01u}$  is upstream ambient Pressure.

The Anti-Surge Control mechanism using PID Controller with different Controllers like Fuzzy logic controller (FLC), Neuro-fuzzy Controller (NFC), and Neutral network Predictive Controller (NNPC) are represented in Fig. 4. In this work, the PID controller is used as the industry standard and other controllers to analyze the anti-surge control operations. The Compressor system produces the mass flow ( $m$ ) and compressor characteristic ( $\psi_c$ ) outputs and inputs the anti-surge mechanism. The different controller's output ( $asc_{in}$ ) is input the recycle value unit. The Surge control line (SCL) is considered along with the surge line (SL), which is linear and horizontal to the surge margin. The compressor characteristic SCL with the mass flow is represented in Equation (13) as follows:

$$\psi_{scl}(m) = im + j \quad (13)$$

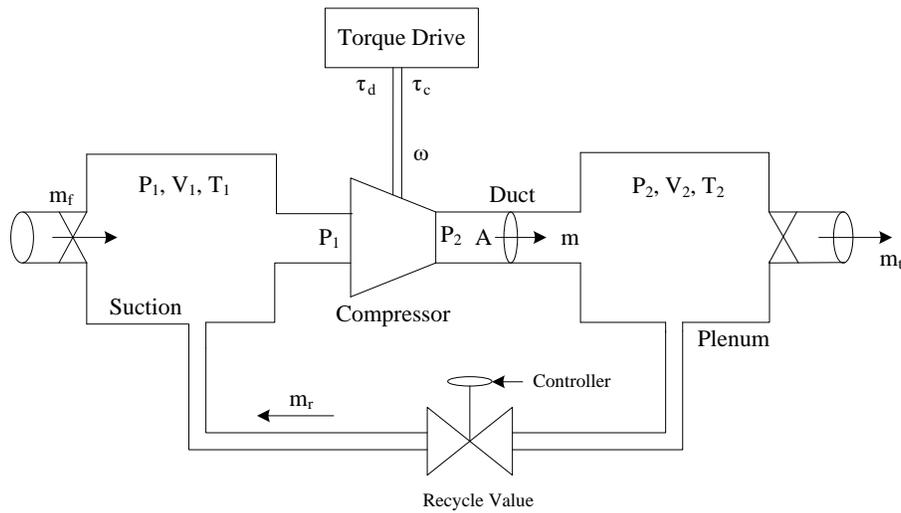


Fig. 2. Compressor Recycle System based Anti-surge Control Mechanism in using different Controllers.

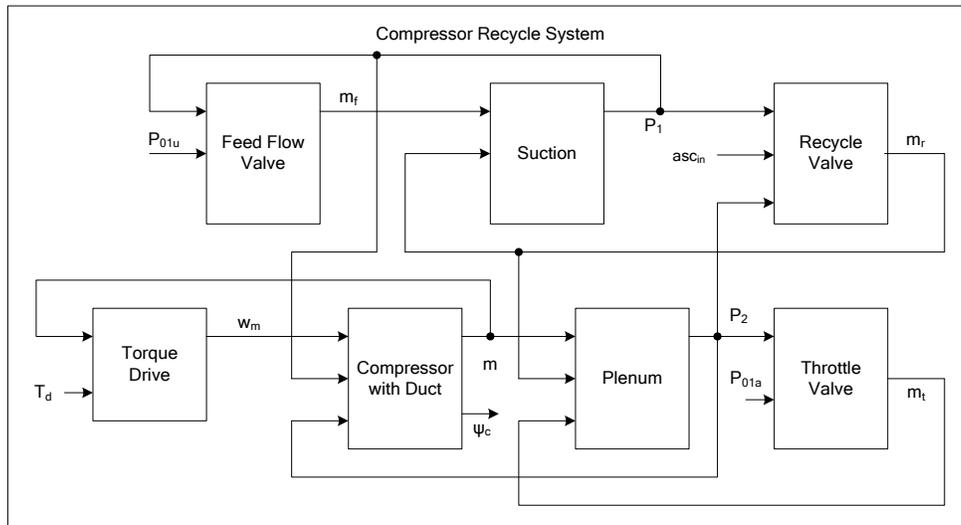


Fig. 3. Modeling of the Compressor Recycle System.

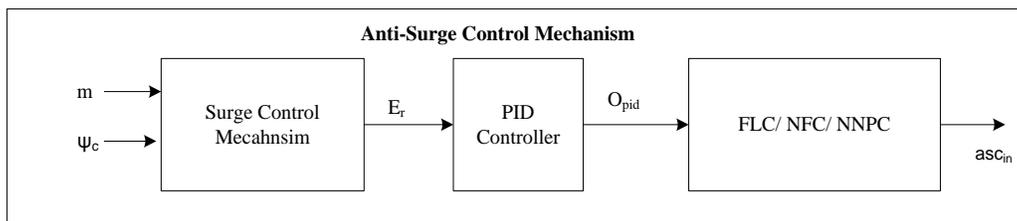


Fig. 4. Anti-Surge Control Mechanism using different Controllers.

Where  $i$  and  $j$  are the coefficients for SCL, when the operating point (OP) is right to the SCL, the control mechanism is not activated.

Only when the operating point (OP) is left to the SCL, the control mechanism is activated. The PID controller and other controllers (FLC/ NFC/NNPC) are used to getting the OP back to the right, with its controlling mechanism. To analyze the controlling mechanism, first, find the distance ( $d_i$ ), as the difference (horizontal) between the SCL and OP, which is represented in Equation (14) as follows:

$$d_i = m - m_{scl}(\psi_c) \quad (14)$$

The OP is calculated using mass ( $m$ ) and  $\psi_c$ . The  $\Psi_{scl}$  ( $m$ ) is same as  $\psi_c$ , when Pressure rises and consider the inverse of Equation (13) to calculate the  $m_{scl}(\psi_c)$ , which is represented in Equation (15) as follows:

$$m_{scl}(\psi_c) = \frac{\psi_c - j}{i} \quad (15)$$

Use Equation (15) to calculate the distance value in Equation (14). The error value is calculated based on the horizontal distance using OP and SCL. When the distance value is +ve, the error value ( $E_r$ ) is set to zero, and no need to perform any controlling operation, and the OP is located to the right position of SCL. When the distance value is -ve, its +ve value is used to calculate the error value ( $E_r$ ) using controlling operation. The error value ( $E_r$ ) calculation is represented in Equation (16) as follows:

$$E_r = \begin{cases} 0 & d_i > 0 \\ -d_i & else \end{cases} \quad (16)$$

This error value ( $E_r$ ) is input to the PID controller and it is represented in Equation (17) as follows:

$$O_{pid} = K_p E_r + K_i \int_0^t E_r dt \quad (17)$$

Where  $K_p$  and  $K_i$  are the proportional gain and integral gain of the PID controller, the PID Controller output ( $O_{pid}$ ) is used directly to control the flow percentage in recycling valve for the anti-surge control mechanism. To improve the anti-surge control with OP, a different controlling mechanism is introduced.

#### A. Fuzzy Logic Controller (FLC)

The Fuzzy logic Controller (FLC) provides a better dynamic response than the conventional PID controllers and is used in most industrial control applications. The FLC mainly contains five blocks: the Compressor recycle system: Fuzzification, De-fuzzification, database (Knowledge base), Ruleset, and evaluation process. The Fuzzification converts absolute (crisp) data into Fuzzy (linguistic) data. The PID controller output is input to the fuzzification process as crisp data and produces the fuzzy data. These fuzzy data values are considered as Fuzzy-sets in FLC. The membership functions analyze these fuzzy sets and estimate whether they are low, high, or oversized. The evaluation process provides the decision to control and study the fuzzy rules stored in the rule base. The FLC process performance is analyzed by rule base. The de-fuzzification process converts back the Fuzzy data into crisp data. The Knowledge-based is used to store the membership functions of both the Fuzzification and de-fuzzification processes. The evaluation of fuzzy rules is done by basic fuzzy sets with AND, OR, and NOT operations. The de-fuzzification output is the final FLC output and is processed in Compressor recycle system. The function of the Fuzzy logic process with the Compressor recycle system is represented in Fig. 5.

The Mamdani Fuzzy interface system (FIS) is used to create the FLC. The two fuzzy variables are used in the FLC: one is PID output ( $O_{pid}$ ), and another is Change in PID output ( $CO_{pid} = O_{pid}(i) - O_{pid}(i-1)$ ). Where  $i$  value is set to 1 to 3.

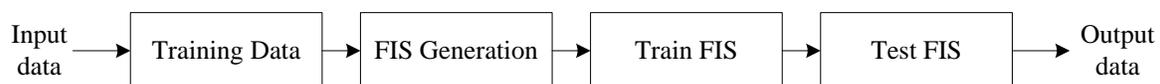


Fig. 6. Operation of NFC.

The two Fuzzy inputs ( $O_{pid}$  and  $CO_{pid}$ ) and FLC output ( $O_{flc}$ ) are normalized to the display range [0 to 0.6] using fuzzy set membership functions for the given universe of discourse. The Fuzzy Input and output use the triangular-shaped membership functions. The Three fuzzy variables are used as  $O_{pid}$  and  $CO_{pid}$ , which includes Negative Big (NB), Zero (ZO), and Positive Big (PB). Human knowledge is used to create the decision logic for fuzzy rule sets. The Fuzzy rule set for Compressor recycle system is tabulated in Table I. The fuzzy variables like NB are in the range of [-0.3 to 0.3], ZO is in the field of [0 to 0.6], and PB is in the range of [0.3 to 0.6]. The nine- rules are used in the fuzzy inputs ( $O_{pid}$  and  $CO_{pid}$ ) with three fuzzy variables. The rule set converts these two fuzzy inputs into single FLC output using the centroid technique in the de-fuzzification process.

#### B. Neuro-Fuzzy Controller (NFC)

The FLC is extended with NFC using an Artificial Neuro-fuzzy interface system (ANFIS). The NFC automatically realizes the FIS using Neural networks (NNs). The NFC provides a conversion mechanism (Crisp to fuzzy and vice-versa) efficiently. The NFC uses the NN to optimize the FIS using the ANFIS system. The operation of the NFC is represented in Fig. 6. The NN has a learning skill set and is capable of training and testing the FIS in NFC. The NFC uses two inputs like PID controller output ( $O_{pid}$ ) and change in PID controller output ( $CO_{pid}$ ) for training. Once the training data is loaded in NFC, generate the FIS using grid partition using the Gaussian membership function. The seven MF's are considered with constant output in NFC.

Next, Train the FIS using the hybrid optimization method. The error tolerance is set to 3 epochs to complete the ANFIS training. Lastly, test the training data to analyze the NFC output.

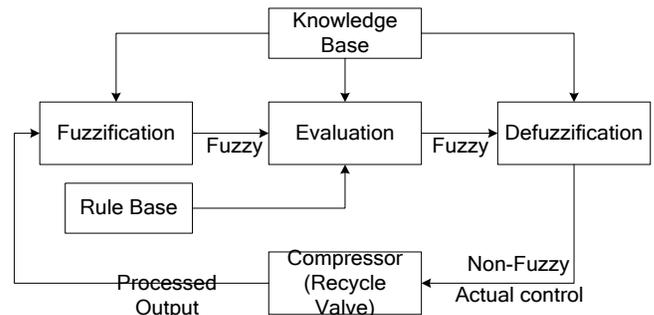


Fig. 5. Fuzzy Logic Process with Compressor Recycle System.

TABLE I. THE FUZZY RULE SET FOR COMPRESSOR RECYCLE SYSTEM

| E/CE | NB | ZO | PB |
|------|----|----|----|
| NB   | NB | NB | ZO |
| ZO   | NB | ZO | PB |
| PB   | ZO | PB | PB |

### C. Neutral Network Predictive Controller (NNPC)

The NNPC mainly contains the NN model, Optimization unit, and Compressor recycle system. The NNPC predicts the future performance of the Compressor recycle system using the NN model. The NNPC Based Compressor recycle system for the anti-surge controlling mechanism is represented in Fig. 7. The NNPC is used to calculate the NN model input ( $m$ ) and further optimize the Compressor recycle system's performance over a defined time horizon.

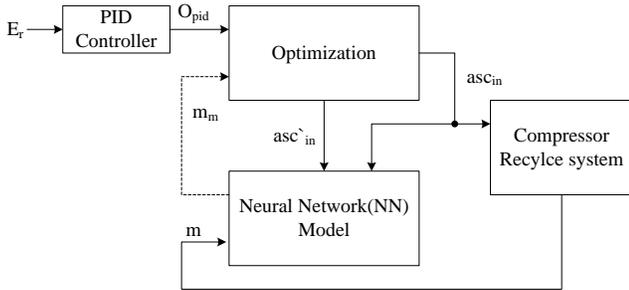


Fig. 7. NNPC based Compressor Recycle System for Anti-surge Control.

The NNPC first identifies the Compressor recycle system (plant) and then uses this model to predict the future performance using a controlling mechanism. In the first stage, the NN model is trained by a predictive control mechanism to analyze the dynamics of the Compressor recycle system. The NN training signal is generated by calculating the predictive error between the NN model output and Compressor recycle system output. The NN model [26] uses the previous inputs (Optimization) and the previous compressor system's output to analyze the Compressor recycle system's output values. In the second stage, the optimization algorithm [27-28] is used for precision and to calculate the control signal. This control signal minimizes the below Equation's (18) performance over a defined time horizon.

$$J = \sum_{j=N_1}^{N_2} (O_{pid}(t+j) - m_m(t+j))^2 + \rho \sum_{j=1}^{N_u} (asc'_{in}(t+j-1) - asc'_{in}(t+j-2))^2 \quad (18)$$

Where  $asc'_{in}$  is a temporary control signal,  $O_{pid}$  is PID controller's output acts as a desired response,  $m_m$  is the response of NN model,  $\rho$  is used to calculate the performance index's sum with squares of control increments.  $N_1$ ,  $N_2$ , and  $N_u$  values define cost and control horizon.

The  $J$  value is minimized using the optimization unit by calculating the  $asc'_{in}$  and inputting the  $asc_{in}$  to the Compressor recycle system. The Recycle valve output is represented after the anti-surge controlling mechanism using NNPC controller is described in Equation (19) as follows:

$$m_r = O_{pid} \cdot asc_{in} \cdot \sqrt{P_2 - P_1} \quad (19)$$

The NNPC uses four input layers, five hidden layers, and 1-output layer for training the data. The 100 training samples with a sample interval of one second are selected for training the data in the NN training tool. The NNPC obtains the six iterations (epoch) with 8.44e-6 performance and six validation checks while training the data.

### V. RESULTS AND DISCUSSION

The Compressor recycle system with an anti-surge control mechanism using a PID controller with different controllers is presented in this section. The Simulation results of NNPC based compressor recycle system are introduced and compared with other controllers like PID controller, FLC, and NFC to realize the performance metrics. The NNPC based Compressor recycle system is simulated with the following assumption. The mass flow ( $m$ ) and compressor speed ( $\omega$ ) is set to zero initially; two volumes of Pressure ( $V_1$  and  $V_2$ ) are defined as ambient. The higher mass flow will be obtained by opening the recycle valve. When more mass flow occurs, the operating point (OP) positions are shifted to the right position in  $\psi c$ . The simulation results of the Compressor recycle system are represented in Fig. 8. The compressor output mass flow ( $m$ ), throttle valve output ( $m_t$ ), recycles valve output ( $m_r$ ), and feed flow valve output ( $m_f$ ) is represented on the left-hand side. The torque drive's speed ( $\omega$ ), suction pressure ( $P_1$ ), and Plenum pressure ( $P_2$ ) are presented on the right side of Fig. 8.

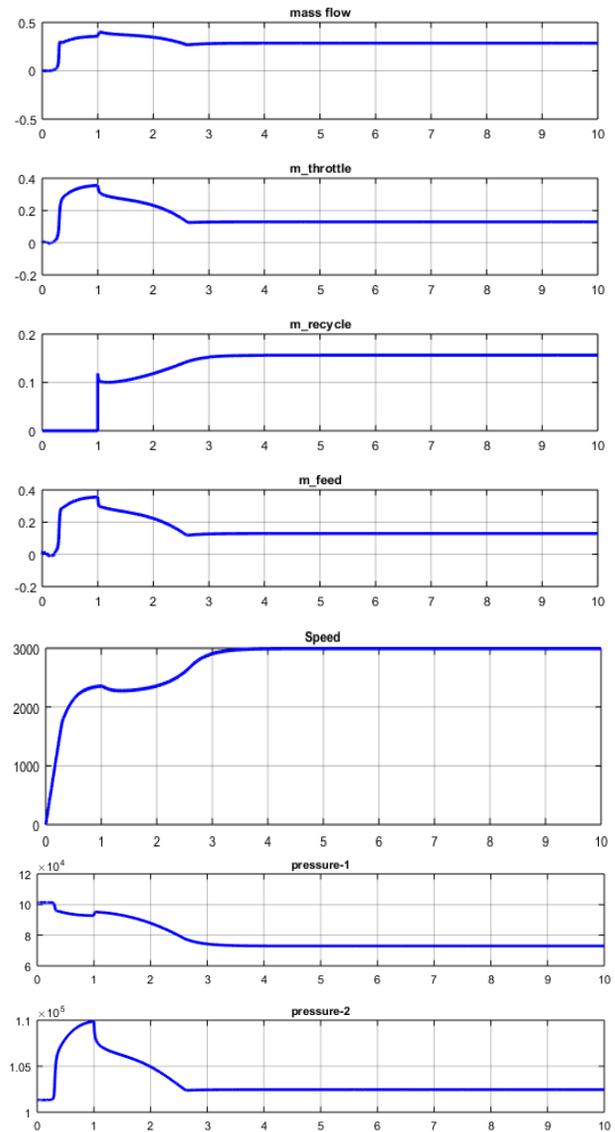


Fig. 8. Simulation Results of Compressor Recycle System.

The mass flow is set to zero initially and varied based on speed, Throttle, and feed flow response. The throttle output (mt) is changed based on plenum pressure ( $P_2$ ) and ambient condition ( $P_{01a}$ ). The recycle output is not activated until the controller responds and varies based on NNPC controlling mechanism. The feed flow ( $m_f$ ) is zero initially and increases to some extent and gradually decreases, which indicates the system enters into surge line (SL). When the feed flow ( $m_f$ ) output is steady, the OP's are shifted from SL to SCL. The torque drive provides 3000 rad/s—speed output ( $\omega$ ) after the controller response. The two pressures ( $P_1$  and  $P_2$ ) outputs are decreased eventually and reach steady-state after NNPC response.

The operating point (OP) position and SL and SCL using different controllers in the Compressor recycle system is represented in Fig. 9. The current OP position is identified using other controllers (PID, FLC, NFC, and NNPC) and

compared with SL and SCL. The SL and SCL are defined initially and perform the simulation of the Compressor recycle system using different controllers to obtain the OP's position.

The SL and SCL are linear and defined with i and j values using Equation (15). The OP is in the left position during the initialization of the Compressor recycle system. Later shifted to the right next to the SCL and moving around it. The OP is very near to SL and turns back to SCL using PID Controller, whereas, in FLC and NFC, The OP is flat and back to SCL. In contrast, The NNPC provides a better OP position, and OP is located next to the SCL. So, the NNPC provides a better OP position than PID, FLC, and NFC in Compressor recycle system.

The different controller's output response for Compressor recycle system is represented in Fig. 10. The output response of all the controllers is initially zero and later varied based on the controlling mechanism in Compressor recycle system.

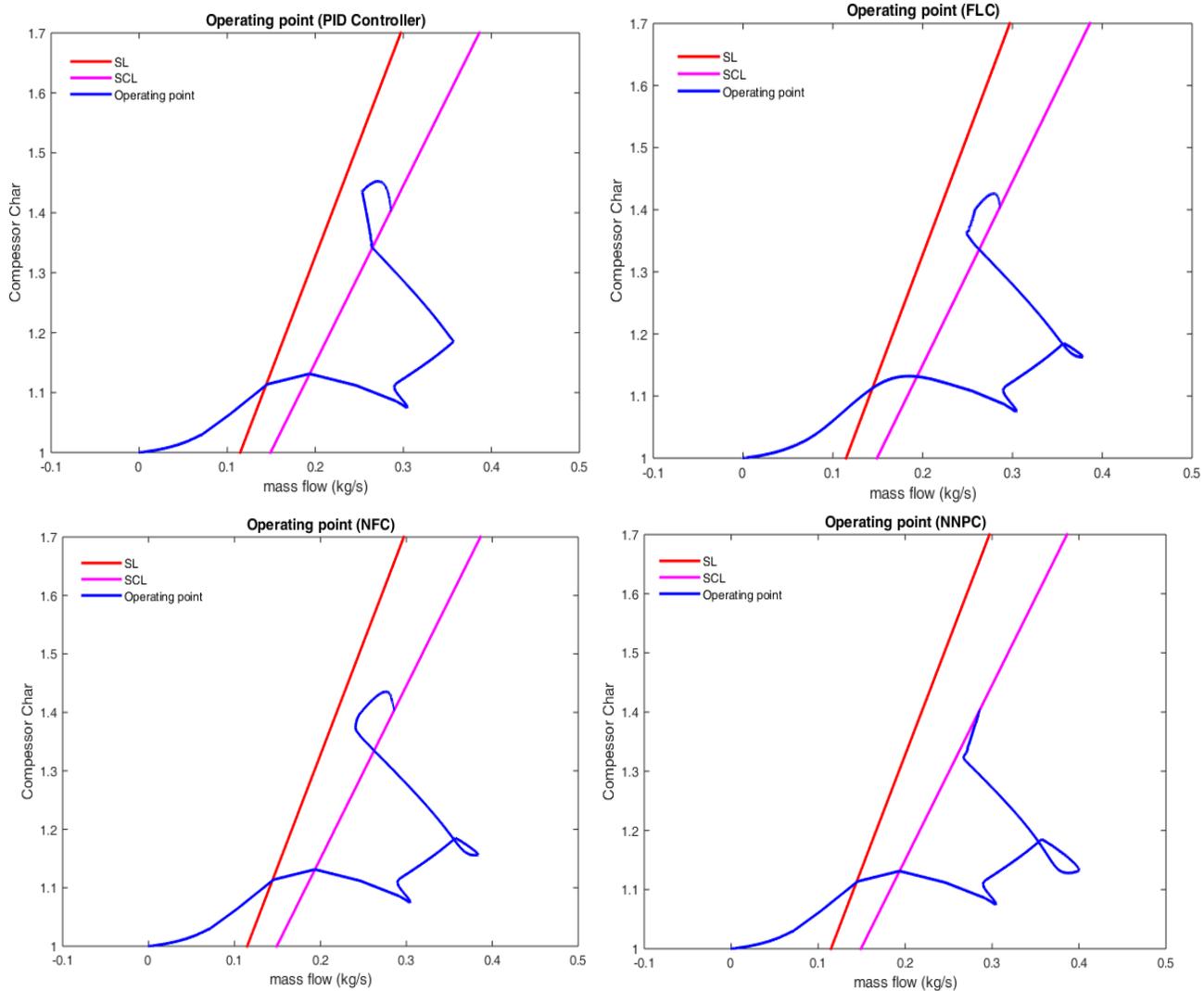


Fig. 9. Different Controller's Operating Points Position in Compressor Recycle System.

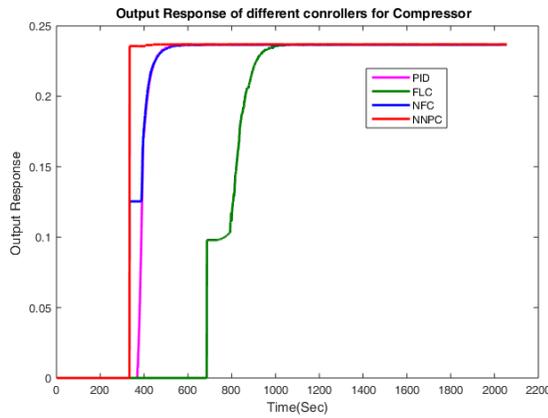


Fig. 10. Dynamic Response of different Controllers for Compressor Recycle System.

The FLC takes more time to provide the output response, whereas NNPC takes less time to provide output response than PID and NFC. The dynamic output response values of different controllers concerning time for the Compressor recycle system are represented in Fig. 11. The output response of the PID controller takes a rise time of 0.976 sec, settled at 4.27 sec with an overshoot (%) of 0.263 sec. Similarly, the FLC controller takes a rise time of 1.905 sec, settled at 3.87 sec with an overshoot (%) time of 0.077 sec. The NFC takes 0.1112 for rising time, settled at 0.236 sec with overshoot (%) time of zero. In contrast, the output response of the NNPC takes a rise time of 0.013 sec, settled at 0.2369 sec with an overshoot (%) of zero. Overall, the NNPC provides a better dynamic output response with less rise time, settled early, and zero overshoot time than PID, FLC, and NFC.

The error analysis of different controllers for the Compressor recycle system is represented in Fig. 12. The error analysis includes Integral Square Error (ISE), Integral Absolute Error (IAE), and integral time absolute error (ITAE) parameters. The error response of the NNPC based Compressor recycle system improves 14.49 % in ISE, 51.9 % in IAE, and 95.6 % in ITAE than the PID-based Compressor recycle system. Similarly, the 4.83 % in ISE, 34.8 % in IAE, and 91.04 % in ITAE are improved compared to FLC. Compared to NFC, the 11.86 % in ISE, 34.92 % in IAE and 95.53 % in ITAE are improved in NNPC based compressor recycle system.

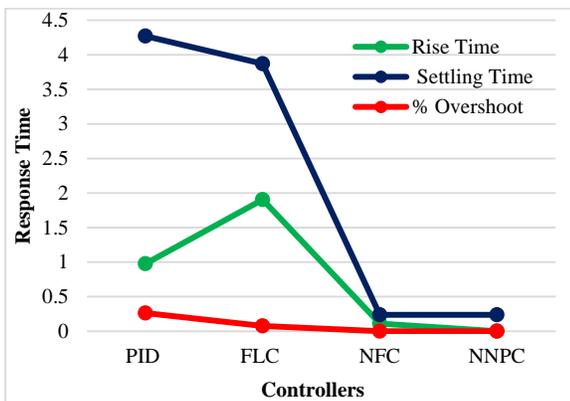


Fig. 11. Output Response Values of different Controllers for Compressor Recycle System.

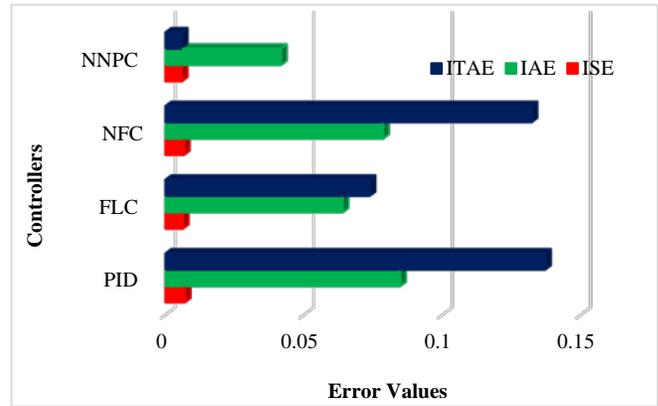


Fig. 12. Error Calculation of different Controllers for Compressors Recycle System.

## VI. CONCLUSION AND FUTURE WORK

In this manuscript, an efficient NNPC based anti-surge control mechanism is modeled for variable-speed centrifugal Compressor recycle system. The NNPC provides higher mass flow in the recycle valve to the compressor system and avoids the surge with better stabilization. The Compressor is working at a more increased mass flow with low Pressure to avoid the surge. The simulation results of NNPC based Compressor recycle systems like mass flow, speed, and Pressure waveforms are discussed. The operating point's position with the Surge line and surge control line for different controllers is represented. The NNPC based mechanism provides dynamic response than PID, FLC, and NFC controller-based mechanisms. The dynamic responses (Rise time, settling time, and % overshoot time) of the NNPC based Compressor recycle system reduces the ISE of 14.49%, 4.83%, and 11.86% in PID, FLC, and NFC based Compressor recycle system. Similarly, NNPC also reduces the IAE and ITAE values than PID, FLC, and NFC-based controller mechanisms. In future, incorporate the proposed work with real-time gas plants to realize the performance metrics. Also extend the recycle system with new approach to analyze the stability of the compressor system.

### REFERENCES

- [1] Gravdahl, Jan Tommy, and Olav Egeland. Compressor surge and rotating stall: Modeling and control. Springer Science & Business Media, 1999.
- [2] Fink, David Alan, Nicholas A. Cumpsty, and Edward M. Greitzer. "Surge dynamics in a free-spool centrifugal compressor system." (1992): 321-332.
- [3] Greitzer, Edward M. "Surge and rotating stall in axial flow compressors—Part II: experimental results and comparison with theory." (1976): 199-211.
- [4] Gravdahl, Jan Tommy, and Olav Egeland. "A Moore-Greitzer axial compressor model with spool dynamics." In Proceedings of the 36th IEEE Conference on Decision and Control, vol. 5, pp. 4714-4719. IEEE, 1997.
- [5] Badmus, O. O., C. N. Nett, and F. J. Schork. "An integrated, full-range surge control/rotating stall avoidance compressor control system." In 1991 American Control Conference, pp. 3173-3180. IEEE, 1991.
- [6] Batson, Brett W. "Invariant coordinate systems for compressor control." In Turbo Expo: Power for Land, Sea, and Air, vol. 78729, p. V001T01A070. American Society of Mechanical Engineers, 1996.
- [7] Gravdahl, Jan Tommy, Olav Egeland, and Svein Ove Vatland. "Drive torque actuation in active surge control of centrifugal compressors." Automatica 38, no. 11 (2002): 1881-1893.

- [8] Bøhagen, Bjørnar, and Jan Tommy Gravdahl. "Active surge control of compression system using drive torque." *Automatica* 44, no. 4 (2008): 1135-1140.
- [9] Shehata, Raef S., Hussein A. Abdullah, and Fayez FG Areed. "Fuzzy logic surge control in constant speed centrifugal compressors." In 2008 Canadian Conference on Electrical and Computer Engineering, pp. 000653-000658. IEEE, 2008.
- [10] Albehigi, Abdulkareem A. Wahab, and Rasha Hyder Hashim. "Experimental and Theoretical Analysis of the Surge in a Centrifugal Compressor." In *Modern Methods of Construction Design*, pp. 317-328. Springer, Cham, 2014.
- [11] Budinis, S., and N. F. Thornhill. "Control of centrifugal compressors via model predictive control for enhanced oil recovery applications." *IFAC-papers online* 48, no. 6 (2015): 9-14.
- [12] Budinis, Sara, and Nina F. Thornhill. "Supercritical fluid recycle for surge control of CO2 centrifugal compressors." *Computers & Chemical Engineering* 91 (2016): 329-342.
- [13] Cortinovis, Andrea, Hans Joachim Ferreau, Daniel Lewandowski, and Mehmet Mercangöz. "Safe and efficient operation of centrifugal compressors using linearized MPC." In 53rd IEEE Conference on Decision and Control, pp. 3982-3987. IEEE, 2014.
- [14] Cortinovis, A., H. J. Ferreau, D. Lewandowski, and M. Mercangöz. "Experimental evaluation of MPC-based anti-surge and process control for electric driven centrifugal gas compressors." *Journal of process control* 34 (2015): 13-25.
- [15] Sheng, Hanlin, Wei Huang, Tianhong Zhang, and Xianghua Huang. "Compressor instability active control via a closed-coupled valve and throttle actuators." *International Journal of Turbo & Jet-Engines* 32, no. 3 (2015): 257-264.
- [16] Gritli, Wafa, Hajer Gharsallaoui, and Mohamed Benrejeb. "PID-type fuzzy scaling factors tuning using genetic algorithm and Simulink design optimization for an electronic throttle valve." In 2016 International Conference on Control, Decision and Information Technologies (CoDIT), pp. 216-221. IEEE, 2016.
- [17] Azeem, Mohammad Fazle, and Abdul Kareem. "A fuzzy logic-based super-twisting sliding mode control scheme for electronic throttle control." In 2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 485-492. IEEE, 2016.
- [18] Saeed, Muhammad Ahsan, Nisar Ahmed, Mujahid Hussain, and Adnan Jafar. "A comparative study of controllers for optimal speed control of the hybrid electric vehicle." In 2016 International Conference on Intelligent Systems Engineering (ICISE), pp. 1-4. IEEE, 2016.
- [19] Mustafa, Mohamed Iqbal Mohamed, Joseph Xavier Rayappan, and Kanakaraj Jagannathan. "A neuro-fuzzy controller for grid-connected heavy-duty gas turbine power plants." *Turkish Journal of Electrical Engineering & Computer Sciences* 25, no. 3 (2017): 2375-2387.
- [20] Liying, Zhang, Xu Ting, Li Xuejun, Li Liyan, Guo Pan, and Zhu Yubao. "Study on Nonlinear Compensation Control Method for Electronic Throttle Valve." In 2018 Chinese Automation Congress (CAC), pp. 1085-1089. IEEE, 2018.
- [21] Taleb Ziabari, Masoud, Mohammad Reza Jahed-Motlagh, Karim Salahshoor, Amin Ramezani, and Ali Moarefianpur. "Robust adaptive control of surge instability in constant speed centrifugal compressors using tube-MPC." *Cogent Engineering* 4, no. 1 (2017): 1339335.
- [22] Abo-khsheem, K. A. "Active Control of Surge Compressor System." *Journal of Electrical & Electronic Systems* 7, no. 3 (2018): 1000267.
- [23] Guan, Xudong, Jin Zhou, Chaowu Jin, Yuanping Xu, and Hengbin Cui. "Influence of different operating conditions on centrifugal compressor surge control with active magnetic bearings." *Engineering Applications of Computational Fluid Mechanics* 13, no. 1 (2019): 824-832.
- [24] ARIBI, Yacine, Razika Zammoum Boushaki, and Hocine Loubar. "Active Surge Control of the Recycle Compression System By Hybrid Adaptive Controller." In 2019 4th International Conference on Power Electronics and their Applications (ICPEA), pp. 1-9. IEEE, 2019.
- [25] Ebrahimi, Seyed Hossain, and Ahmad Afshari. "An Artificial Neural Network Model for Prediction of the Operational Parameters of Centrifugal Compressors: An Alternative Comparison Method for Regression." *Journal of Sciences, Islamic Republic of Iran* 31, no. 3 (2020): 259-275.
- [26] Asgari, Hamid, XiaoQi Chen, Mohammad B. Menhaj, and Raazesh Sainudiin. "Artificial neural network-based system identification for a single-shaft gas turbine." *Journal of Engineering for Gas Turbines and Power* 135, no. 9 (2013).
- [27] Zhong, Lulu, Yang Liu, Jun Zhao, and Wei Wang. "Deep predictive controller designed for centrifugal compressor system anti-surge." In 2020 Chinese Automation Congress (CAC), pp. 6554-6559. IEEE, 2020.
- [28] Shestopalov, Mikhail Yu, Ruslan I. Smirnov, and Damir H. Imaev. "Approximation of the Natural Gas Pumping Compressor Characteristics using a Multi-layer Neural Network." In 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), pp. 1088-1091. IEEE, 2021.

# Design and Implementation of True Parallelism Quad-Engine Cybersecurity Architecture on FPGA

Nada Qaim Mohammed<sup>1</sup>, Amiza Amir<sup>2</sup>, Muataz Hamed Salih<sup>3</sup>, Badlishah Ahmad<sup>4</sup>

Faculty of Electronic Engineering Technology, Universiti Malaysia Perlis (UniMAP), Ulu Pauh, Perlis, Malaysia<sup>1,2,4</sup>  
IR4.0 and Intelligent Automation Group, Design and Engineering, Flex, Penang, Malaysia<sup>3</sup>

**Abstract**—Applications, such as Internet of Things, deal with huge amount of transmitted, processed and stored images that required a high computing capability. Therefore, there is a need a computing architecture that contribute in increasing the throughput by exploiting modern technologies in both spatial and temporal parallelisms. This paper conducts a parallel quad-engine cybersecurity architecture with new configuration to increase the throughput. using DE1-SoC and Neek FPGA boards and HDL. In this architecture, each engine operates with 600MHz maximum frequency. Each image is divided into four parts of equal size and each part processed by single engine concurrently to achieve spatial parallelism. Internally, engine is handling image's part in temporal parallelism and deep pipelining abstraction applied in every engine by dividing it to sub modules to execute different tasks concurrently. All data processed in engines is encrypted via AES algorithm that implemented as a significant part of engine architecture. The obtained results increased the throughput by four times, with 153,600Mbps, that make this computing architecture efficient and suitable for fast applications such as IoT and cybersecurity level of processing.

**Keywords**—Field programmable gate array (FPGA); spatial parallelism; cybersecurity; throughput component

## I. INTRODUCTION

Data processing and transfer and store have grown dramatically in the last year because they are used in different applications via different communication networks. This growth requires an exploitation of modern technologies to enhance and increase this processing in a parallelism manner to achieve high throughput. Various algorithms and devices introduced by researchers to achieve this goal in both software and hardware implementations by exploit the available modern techniques and possibilities in temporal and spatial parallel processing [1]. Gaining high productivity requires devices that are reconfigurable, work in real-time, efficient, low power consumption, and the ability to perform parallel processing features. One of the candidates' approaches to spatial parallel processing is field-programmable gate arrays (FPGAs), whose features meet this. So, one can take advantages of FPGAs to implement a platform to achieve efficient parallelism for real time processing applications [2].

At the same time, much of the data, which are transferred through cyberspace, include confidential or personal information, so it has become a target for hackers and adversaries to access, change, or damage them. Therefore, it is necessary to use techniques that maintain the confidentiality and integrity of these data. One efficient way to achieve this is

cryptography, in which various algorithms that vary in their security are proposed. An algorithm of cryptography, which has not yet been found a way to break it, is the AES algorithm [3]. However, the AES algorithm has high computational power to achieve its operations because of the large number of rounds that are used, and the need for more time to encrypt the data [4,5].

To obtain faster and more efficient computation power to encrypt massive data with high throughput, the AES algorithm must be implemented in a parallel manner. Therefore, the FPGA is a candidate approach to hardware implementation of the AES algorithm to ensure data protection, high-speed encryption rate, and high throughput.

This paper focuses on hardware architecture implementation on FPGA that based on true spatial and temporal parallelism using quad engines for cybersecurity. This architecture works through partition each image into four parts and distributes these parts to multiple engines that can process data in temporal and spatial parallelism the image parts concurrently.

The paper is organized to include a survey related work that has been conducted by some academic and researchers in Section II. Section III includes a detail description of the proposed architecture. Section IV contains the implementation results. It includes a new architecture to improve the performance in terms of the operating frequency and throughput of the AES algorithm. The results of the simulation and a comparison with previous studies are presented in Section V. Section VI presents the conclusions of the study.

## II. RELATED WORK

Different approaches have been proposed to enhance hardware implementation to obtain more image processing performance. The widespread deployment of field-programmable gate arrays (FPGAs) has enabled multi-processing in real-time processing applications, which has accelerated massive spatial parallel applications. Throughput increasing is one of the important criteria in measuring the efficiency of the used technique. To achieve this, there is a need to make use of spatial parallelism and duplicate processing units that execute simultaneously. This can done by divide the main task into several subtasks, each subtask is executed one processing elements [5]. Combining the FPGA features with spatial parallelism will help to decrease the complexity, cost, and power consumption and increase the throughput of the proposed systems. In [6], a design and

simulation was proposed to enhance the images using VHDL language using Xilinx Virtex- 2 Pro FPGA and MATLAB. In [7], the AES algorithm was implemented using Xilinx's Virtex-E and Virtex-II devices to achieve faster FPGA – based implementation. The obtained experimental results were throughput of 17.80 Gbps on a Virtex-II with a clock frequency of 139.1 MHz and 10750 slices were used.

The Altera FLEX FPGA family utility was used in [8] to execute different types of algorithms that deal with the image as a whole image instead of dealing with pixels values individually. For image encryption, Chang et al. al in 2009 implemented the AES algorithm on an FPGA using a Virtex22 device. The core of the AES was 32 bits, and it occupied 104 slices. The throughput of implementation was 794 Mbps and the efficiency was 7.93 Mb/slice [9]. In 2010, Kumar and Purohit implemented a 128-bit AES algorithm on an FPGA using a Xilinx Spartan 3 device to achieve high speed using low-cost devices [10].

In [11], Manoj and Manjula implemented an AES 128 bit using a Xilinx Spartan 6 device, 8-bit input (data pixels), and unrolled them to 128 bits. The throughput of encryption was 252.132 Mb/s, and the efficiency was 0.53 Mb/slice. In [12], Karimian, Rashidi, and Farmani implemented a 128-bit AES algorithm using an Altea Stratix device, achieving a throughput of 617 MB/s and an efficiency of 0.76 Mb/slice. In [13], the throughput obtained from implemented a 128-bit AES using a Xilinx Spartan was 3.40 GB/s and efficiency of 5.43 Mb/slice. In [14], some algorithms for image enhancement were implemented using the Spartan3E FPGA kit to obtain high-performance digital signal processing applications. MATLAB, Xilinx ISE, Verilog HDL, and ModelSim were used for this implementation. In [15], Xilinx 14.2 software for 2D and 3D image enhancement was used to design and develop algorithms to enhance the size of image pixels. In [16], the image enhancement and DE noising process was introduced using point processing methods by using a partial dynamic reconfiguration of the FPGA to decrease the requirement of resources and increase the performance. In [17], image enhancement approaches were introduced to enhance grayscale images by adopting a custom hardware processor on an FPGA. The implementation depended on the use of filters and the VERILOG hardware description language. The achieved image enhancement of these approaches used neighborhood processing operations in the spatial domain using parallel processing.

In [18], Rahimunnisa et al. the proposed structure is implemented in a Virtex-6 XC6VLX75T FPGA device, which gave a throughput of 37.1 Gb/s with a maximum frequency of 505.5 MHz in [19], Groth implemented AES encryption on a Xilinx Kintex27 FPGA for application data of the biometric image. The implementation throughput was 40 GB/s with an efficiency of 5.27 Mb/slice and a powerful performance of 286 GB/w. In [20], techniques to perform task-level out-of-order execution were proposed and implemented using the Xilinx Virtex-5 FPGA device to improve flexibility. The implementation results showed a better efficiency in terms of performance and resource usage. In [21], AES was implemented on FPGA; the obtained experimental results were a throughput of 113.5 GB/s on a Spartan-6 device. In [22], the

focus was on exploring the features of FPGA parallelism to process image applications in spatial parallelism for real-time image processing. They used board DE2-115 as a vehicle project, while the VHDL hardware description language was used as the hardware design of the system. The operating frequency was 1GHz, and the system could be reconfigured to implement several algorithms using the same hardware. In [23]. In [24] architecture was conducted to optimize parallel processing and implement this architecture on FPGA. It improved the consumption of the power by 90% in compared with others. Its throughput was 1.34 GB/s at a 131.16 MHz operating frequency for processing images with a size of  $512 \times 512$ .

Presented two pipelined algorithms for effective processing time reduction using pipelined and parallel techniques to process the AES encryption algorithm using FPGA. Xilinx "Spartan-3A/3AN FPGA Starter Kit was used in the implementation of algorithms. The results showed that parallel implementation was good.

### III. PROPOSED HARDWARE PARALLELISM QUAD-ENGINE ARCHITECTURE

The introduced architecture aims increasing the operating frequency and throughput, while decrease the power dissipation, area, and latency by exploiting a quad-engine with true parallelism. Each engine operates with 64 bits and a frequency of 600 MHz.

This section provides the structure of the proposed architecture and the mechanism for performing true parallelism quad-engine cybersecurity. The architecture in the spatial domain begins with the top-level design of the proposed architecture, as shown in Fig. 1.

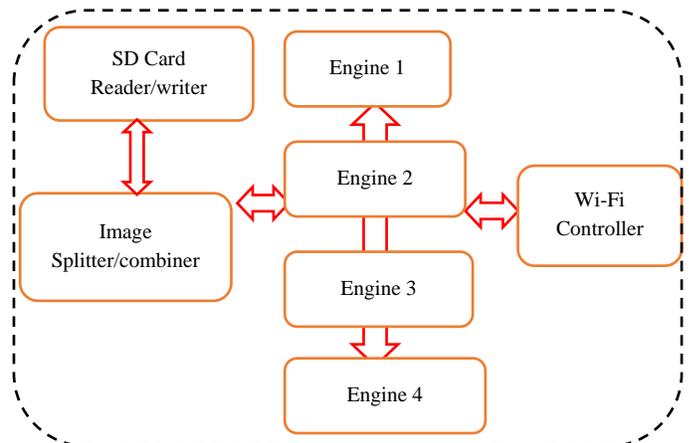


Fig. 1. System Top Level Design and Implementation True Parallelism Quad-Engine Cybersecurity Architecture.

#### A. Hardware Components

To implement the proposed approach, the following components are used:

- SD Card: used Micro SD card interface to read and store images.
- FPGA board: DE1-SoC board and NEEK board with the same TLD are used.

- Wi-Fi controller: Wi-Fi controller in DE1 as (Server) and Wi-Fi controller in NEEK board as (Client) to transfer the image between the two boards.

### B. Software Component

The VHDL, synthesized by Altera Quartus II and simulated using Modelsim are used to implement the architecture.

### C. The Proposed Algorithm

The design steps of the algorithm take traditional AES and apply it to an FPGA platform to exploit spatial and temporal parallelism to increase throughput of cybersecurity encryption/decryption processing.

Input: Images need encryption

Output: Encrypted message.

- Read four images each time and stores them on an SD card.
- Split each image spatially into equal four parts. Each part is represented by a matrix and takes a number of two digits, where the first digit is part number and the second digit is image number, (for image 1, the parts number are 11, 21, 31 and 41), Fig. 2.
- Make cycle shifting to the image parts. The shifting amount depends on the image number, as in Fig. 3.
- Sent each one of the four parts independently to one of the four-engine in the encryption/decryption unit, each engine will encrypt 1/4 image using AES 128-bit block.
- In each engine, a deep pipelining is used by divide it into sub modules to process different task.
- After encrypt the image parts, they are merged to create the encrypted image.

The cycle shifting of image parts is done as follow:

- Engine 1, will send the first part of all 4 images (11, 12, 13 and 14).
- Engine 2, will rotate (as a ring counter) forth index in this image 2 to become the first part (24, 21, 22 and 23), cycle shift by 1.
- Engine 3, will rotate again to start with the next index in image 3, cycle shift by 2.
- Engine 4, will rotate again to start with the next index in image 4, cycle shift by 3.

To decrypt the image, the encryption steps are executing in reverse order.

The architecture processes four parts at the same time, using spatial and temporal parallelism, where each engine processes one part from each image with 600MHz and 64 bits. The input images used in this approach can have different colors or grays of any size. The WIFI controller controls the transfer between the DE1-SoC board and NEEK board. It is necessary for the two boards to have the same TLD.

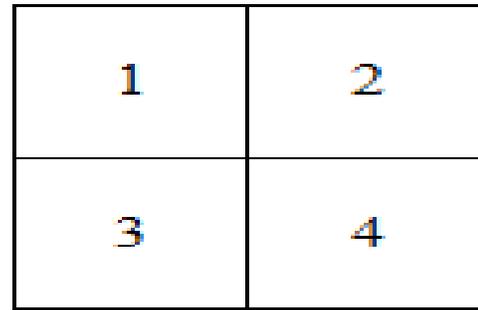


Fig. 2. Quad Sub Image.

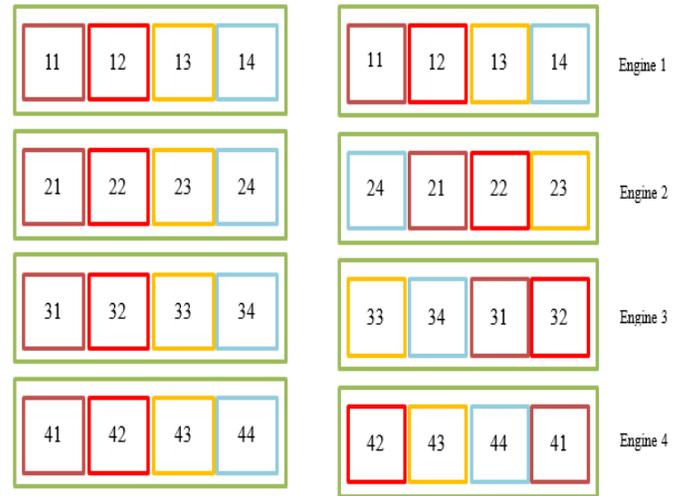


Fig. 3. Shifting Sub Image.

The throughput, number of bits processed per unit time, is measured using the following equations, specified in Mbps or Gbps.

$$\text{Throughput} = (F_{\max} * 4 * \text{No. of bits processed}) \quad (1)$$

Where  $F_{\max}$  represent the maximum frequency and in the proposed design, taking 600 MHz for  $F_{\max}$ , No. of bits equals to 64 bit, whereas, the 4 represents the number of used engine.

## IV. IMPLEMENTATION

The proposed architecture was used to implement the traditional AES algorithm as an application to encrypt four image at the same time.. To encrypt these images the DE1\_Soc, Neek board FPGA device, and Wi-Fi (server and client) are connected together. The Altera Quartus Prime18.1 tool used for the synthesis. These images may have different types of colors or gray of any size.

First, the images were read, as shown in Fig. 4. After reading the image will be doing splitting process, each image was split into four sub-images. Fig. 5 illustrates the sequence of quad sub image 1, 2, 4, and 5.

Then perform the cycle shifting for parts of each image independently, Fig. 5.

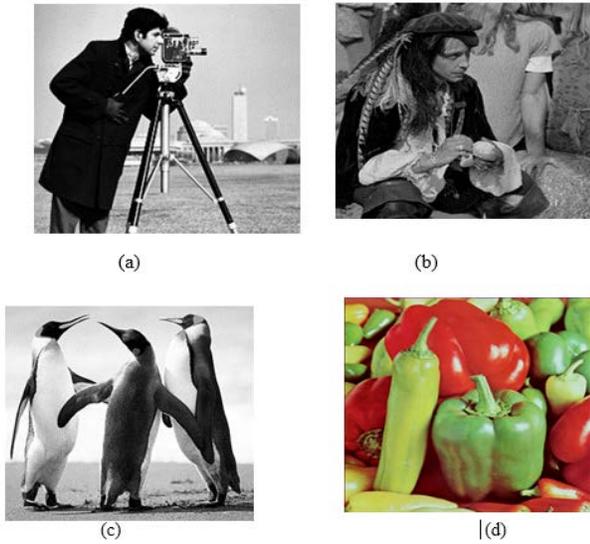


Fig. 4. Input Image.

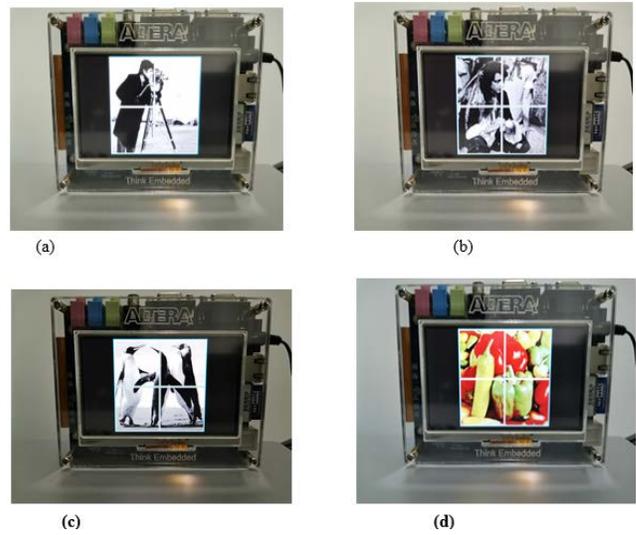


Fig. 6. Split each Image into 4 Sub-Images.



Fig. 5. Step 1 Split Image to Sub Images.



Fig. 7. Split Image Step to Sub Images.

Now, each part is transferred to a separate engine for execution, which performed in a pipelining manner, temporal parallelism. The engine consists of the necessary operation of an encryption/decryption unit. Each engine operates at 600 MHz, and each sub-image part is encrypted by one engine. The four sub-image encryption was performed simultaneously, as shown in Fig. 6.

In Fig. 6 to 10 shows images of real-time execution on FPGA that are highlighted on the LCD touch screen of the NEEK board. VHDL is used to write the design code, in addition to use Altera Quartus II for synthesized. Fig. 6 shows the sub-images of the four image parts after the splitting process is performed.

Each part of the image is sent independently to one of engines, and then a new part of new image following it. Sub images are reordered and sent to make the system more secure.



Fig. 8. Real Live Implementation of First Splitting on NEEK Board.

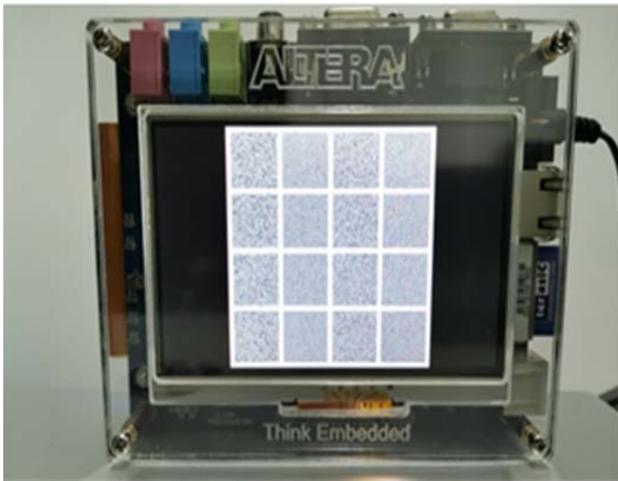


Fig. 9. Real Live Encryption of Images Splitting on NEEK Board.



Fig. 10. Real Live Implementation of Second Splitting on NEEK Board.

Table I includes the details Altera Quartus utilization and Table II includes performance comparisons with other works.

TABLE I. SUMMARY OF DEVICE UTILIZATION USING ALTERA QUARTUS

| Resources           | Available                | Used    | Utilization |
|---------------------|--------------------------|---------|-------------|
| Total Logic Element | 25K logic elements (LEs) | 19,351  | 77.4%       |
| Dedicated Register  | 50                       | 43      | 86%         |
| Memory Bits         | 66 M9K                   | 428,133 | 68%         |

TABLE II. PERFORMANCE COMPARISON

| Reference        | FPGA Device             | Throughput   |
|------------------|-------------------------|--------------|
| Our architecture | The DE1_Soc, Neek board | 153,600Mbps  |
| [7]              | Virtex-II               | 17.8 Gbps    |
| [9]              | Virtex 22               | 794 Mbps     |
| [11]             | Xilinx Spartan 6        | 252.132 Mbps |
| [12]             | Altera Statix           | 617 Mbps     |
| [18]             | Virtex 6                | 37.1 Gbps    |
| [21]             | Spartan 6               | 113 Gbps     |
| [23]             | Spartan 3A/3AN          | 1.34 Gbps    |

The obtained results showed the increasing in the throughput by four times, the throughput rate becomes 153,600Mbps that make this introduced architecture efficient and suitable for fast applications Area utilization was specified with respect to the number of slices used in Altera FPGAs. Additionally, four look-up tables (LUTs) and eight storage elements exist in each slice of the Altera FPGA.

## V. CONCLUSION

This study demonstrates the architecture ability to perform the AES algorithm in spatial and temporal parallelisms. To implement this project, several Altera® Nios II Embedded Evaluation Kit and Cyclone III Edition features were harnessed, such as switch inputs and the LCD touch screen alongside the LEDs. Each of the four engines operates with maximum 600 MHz clock frequency, and the implementation results throughput rate is 153,600Mbps. These results make this introduced architecture efficient and suitable for fast image processing, such as using complex cyber security algorithms for secure information.

## ACKNOWLEDGMENT

I would like to thank my supervisors Dr.Amiza Amir and Dr. Muataz Hameed, UniMAP staff, my family, and everyone who supported me to make this work.

## REFERENCES

- [1] A.A Purkayastha, S.A. Shidhibhavi, and H.Tabkhi , “Taxonomy of Spatial parallelism on FPGAs for Massively Parallel Applications”,in proc 31st IEEE International System-on-Chip Conference (SOCC), Arlington, VA, USA). pp. 55-60. . (2019).
- [2] N. Q. Mohammed , M. H. Salih , R. Aliana , Q. M. Hussein and N. and Aldeen A. Khalid, “FPGA Implementation of Multiple Processing algorithms using spatial parallelism” , ARPN Journal of Engineering and Applied Sciences, VOL. 13, NO. 15, PP.4556-456., 2016.
- [3] N. Q. Mohammed, Q. M., Hussein, S. M. A,K, and Layth A.A, “Hybrid Approach to Design Key Generator of Cryptosystem”, Journal of Computational and Theoretical Nanoscience, Volume 16, Number 3, pp. 971-977, 2019.
- [4] M. E. Hameed, M. M. Ibrahim, and N. A. Manap, “Review on Improvement of Advanced Encryption Standard (AES) Algorithm based on Time Execution, Differential Cryptanalysis and Level of Security”, Journal of Telecommunication, Electronic and Computer Engineering, Vol. 10, No. 1, pp. 139 – 145, 2008.
- [5] F. A. Habib and Q. M. Hussien, “Survey on Data Security Techniques in Internet of Things” , AL-Kunooze Scientific Journal, vol. 2, no. 2, pp: 27 – 37, 2021.
- [6] L. Lan, “The AES Encryption and Decryption Realization Based on FPGA”, Seventh International Conference on Computational Intelligence and Security, Sanya, China, pp 603-607, 2011.
- [7] K. Jarvinen, M. Tommiska, and J. Skytta, “A Fully Pipelined Memoryless 17.8 Gbps AES-128 Encryptor,” Proceedings of the 2003 ACM/SIGDA Eleventh International Symposium on Field Programmable Gate Arrays, Monterey, California, USA pp 207-215.
- [8] O. F. Yousif, M. H Salih, L. A. Hassnawi, M. A. Albreem, M. Q Seddeq, and H. M Isam, "Design and implementation computing unit for laser jamming system using spatial parallelism on FPGA", In proc. IEEE International Conference on Signal and Image Processing Applications (ICSIPA), Kuala Lumpur, Malaysia, pp. 38-43, 2015.
- [9] K.H. Chang, Y.C. Chen, C.C. Hsieh, C.W. Huang and C.J. Chang, “Embedded a low area 322bit AES for image encryption/decryption application” in Proc. IEEE International Symposium on Circuits and Systems, Seoul, pp. 1922–1925, 2009.
- [10] Y. Kumar and P. Purohit, “Hardware Implementation of Advanced Encryption Standard” in Proc. International Conference on

- Computational Intelligence and Communication Networks, Bhopal, pp.4402442, 2010.
- [11] B. Manoj and N. Manjula, "Image Encryption and Decryption using AES. International," Journal of Engineering and Advanced Technology (IJEAT), vol. 1, no. 5, pp. 210822112, 2012.
- [12] G.H Karimian., B. Rashidi. and A. Farmani, "A High Speed and Low Power Image Encryption with 1282Bit AES Algorithm," International Journal of Computer & Electrical Engineering, vol. 4, no. 3, pp. 290-294, 2012.
- [13] M. Gore. and V. Deotare, "FPGA Implementation of Area Optimized AES for Image Encryption/Decryption Process", international journal of next generation computer application (IJNGCA), vol. 1, no. 9, pp 23-26, 2013.
- [14] P. vanaparthi., G. K. Sree and C.D. Naidu. "FPGA implementation of image enhancement algorithms for biomedical image processing", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. Vol. 2, no. 11, pp. 5747-5753, 2013.
- [15] P.S., R. India, A.Kumar and N.Singh, "FPGA Implementation of 2D and 3D Image Enhancement Chip in HDL Environment", International Journal of Computer Applications, vol. 62, no. 21, pp. (0975-8887), 2013.
- [16] C. K. Sundaram, M. Elango and P. Marichamy, "Implementation of Image Processing Algorithm Using Partial Dynamic Reconfiguration in FPGA", International Journal of Innovative Research in Science, Engineering and Technology. Vol. 3, no. 3, pp. 1457-1462, 2014.
- [17] K. B. Ravi Teja, A. S. Warriar, A. S. Belvadi, and D. R. Gawhane, "Design and Implementation of Neighborhood Processing Operations on FPGA using Verilog HDL", IOSR Journal of VLSI and Signal Processing (IOSR-JVSP). Vol. 4, no. 1, pp. 75-80, 2014.
- [18] K. Rahimunnisa, P. Karthigaikumar, Soumiya Rasheed , J. Jayakumar and S. SureshKumar, "FPGA implementation of AES algorithm for high throughput using folded parallel architecture, security and communication network, vol. 7, pp. 2225-2236, 2014.
- [19] T. H. Groth, "FPGA Optimization of Advanced Encryption Standard Algorithm for Biometric Images", M.S. thesis, Luleå University of Technology, Sweden, 2014.
- [20] Chao Wang, Junneng Zhang, Xi Li, Member, Aili Wang, and Xuehai Zhou, "Hardware Implementation on FPGA for Task-Level Parallel Dataflow Execution Engine", IEEE transaction on parallel and distributed systems, vol. 27, no. 8, pp. 2303-2315, 2016.
- [21] U. Farooq and M. F. Aslam, "Comparative analysis of different AES implementation techniques for efficient resource usage and better performance of an FPGA," Journal of King Saud University-Computer and Information Sciences, vol. 29, no. 3, pp. 295–302, 2017.
- [22] N.Q. Mohammed , M.H. Salih, R. Aliana, Q. M. Hussein and N. A. Khalid, "Design and implementation Image Processing functional units using spatial parallelism on FPGA", ARPN Journal of Engineering and Applied Sciences, vol. 13, no. 15, PP. 4514-452, 2018.
- [23] A. Phadikar, H. Mandal and T. L. Chinu, "Parallel hardware implementation of data hiding scheme for quality access control of gray scale image based on FPGA", Multidimensional system and signal processing, volume 31, pp. 73-101, 2020.
- [24] M.Nabil1 , A.A. M. Khalaf2 , S.M. Hassan Design and implementation of pipelined and parallel AES encryption systems using FPGA. Indonesian Journal of Electrical Engineering and Computer Science, Vol. 20, No. 1, pp. 287-299. 2020.

# Cotton Crop Yield Prediction using Data Mining Technique

Amiksha Ashok Patel<sup>1</sup>

Research Scholar, Faculty of Computer Science and Applications, Charotar University of Science and Technology, Changa, Gujarat, India

Dr. Dhaval Kathiriya<sup>2</sup>

Director Information Technology  
Anand Agricultural University  
Anand, Gujarat, India

**Abstract**—Cotton is a very important crop, as India leads it in terms of production in the world; and also that a vast number of manpower is engaged in farming as well as post-harvest processing and management of different derivatives of it. Weather is crucial for the productivity of the crop. The challenges of climate change; availability of limited land and water for farming; lack of knowledge for good cultivation practices and judicious use of agricultural inputs with farmers are critical hindrances for improving productivity. This requires thorough research on land preparation and use, how to improve fertility of soil, good agronomic practices in lieu of variable climatic conditions, etc. All the talukas of the three districts of North Gujarat where cotton is cultivated have been selected purposively for this study. The effect of soil type, soil pH, soil organic carbon, phosphorous, potassium, precipitation and temperature were selected as independent factors. The yield of cotton crop has positive correlation with the selected parameters. The data sets were applied for analytical process to WEKA. The difference between average of predicted and actual yields of all talukas for high rainfall year 2013 was only 1.55 per cent. The difference between actual and predicted yield for the low temperature year (2015) in different talukas of all talukas was only 0.44 per cent.

**Keywords**—Data mining; cotton crop yield prediction; agriculture; data processing; data visualization

## I. INTRODUCTION

One of the great challenges of agricultural development is to guarantee food security. Simultaneously, securing fiber requirement is also one of the necessities of human being. Cotton is an important crop for world's poor. Cotton is grown commercially in more than 80 different countries, mostly in the longitudinal band between 37°N and 32°S.

Cotton is especially adapted to semi-arid and arid environments, where it is either grown as rain-fed or with irrigation. About 53 per cent of the world's cotton growth areas benefit from full or supplementary irrigation. Cotton has certain resilience to high temperatures and drought due to its vertical tap root. The crop is, however, sensitive to water availability, particularly at the stage of flowering and boll formation. Rising temperatures favor development of the cotton plant, unless day temperatures exceed 32° C.

Climate change will affect the cotton crop in numerous ways in different areas. Temperatures are expected to increase all over India. Rainfall intensity during monsoons may become a prevalent problem. Higher temperatures in already hot areas

may hinder cotton development and fruit formation. Rain-fed cotton production may suffer from higher climate variability leading to periods of drought or flooding. With respect to the production level, cotton has limited capacity to respond to heat stress, through 'compensatory growth'. Its vertical tap root also provides resilience against spells of drought, but also makes it vulnerable to water-logging.

Cotton is a natural plant fiber which grows around the seed of the cotton plant. Fibers are used in the textile industry. First, the cotton fiber is obtained from the cotton plant and then spun into yarn. Further, the cotton yarn is woven or knitted into fabric. The use of cotton has a long tradition in the clothing industry due to its desirable characteristics. The value of world cotton production in 2017-18, was around US\$50 billion. Cotton is a driver of economic development and is of critical importance to the economies of developing and least developed countries. Cotton connects people to markets and provides economic opportunities on the frontiers of the world economy.

Of course, there are many factors that affect prediction of cotton. However, crop yield prediction is extremely challenging due to numerous complex factors which affect cotton crop at different growth stages (a short list is as per Annexure 1). It is very difficult to have a site specific measurement of each of these factors and to evaluate their combined effect on cotton production. As such, the current study was planned to look at the effect of the key seven parameters and their contribution to production.

## II. BACKGROUND

Cotton is one of the major cash crops grown in India. The productivity of this crop can be improved dramatically if correct agro-technologies are adhered. The yield gap of research farms or potential of a variety and that of average harvest at farmers' fields is very high.

Cotton is a crucial component of the Indian economy as her textile industry is predominantly cotton based. India is one of the biggest producers and also exporter of cotton fabric. Cotton cultivation is a well-established practice in India. Gujarat, Maharashtra, Telangana, Andhra Pradesh and Karnataka are the major cotton producer states in India. Indian textile industry contributes to around 5 percent to the nation's gross domestic product (GDP), 14 percent to industrial production and 11 percent to total export earnings. After agriculture, textile industry is the second largest employer of over 510 lakh people directly and 680 lakh people indirectly.

The challenges of climate change; availability of limited land and water for farming; lack of knowledge for good cultivation practices and judicious use of agricultural inputs with farmers are critical hindrances for improving productivity. This requires thorough research on land preparation and use, how to improve fertility of soil, good agronomic practices in lieu of variable climatic conditions, etc. The analytical issues, till yesterday, were been handled by applying different statistical tools. However, using tools of data mining and other diagnostic approaches are becoming more useful in making decisions regarding production practices and also prediction of yields [4].

The comprehensive approach, which comprises of various technologies and methods, for example, statistics, Data Mining, Visual Data Mining (VDM), information handling, Data Warehousing (DW), Online Analytical Processing (OLAP) and different frameworks, was considered to be a useful approach. It can also help in better crop predictions based on historical data.

#### A. Role of DM in Agriculture

An accurate estimate of crop production and risk helps the country in planning supply chain decision like production scheduling. Business such as seeds, fertilizers, agrochemicals and agricultural machinery industries plan production and marketing activities based on crop production estimates [9], [14]. These are helpful for the farmers and the government in decision making namely:

- 1) It helps farmers in providing the crop yield record with a forecast, so as to reduce the risk of crop management.
- 2) It helps the government in making policies for crop insurance and supply chain operations.

In large data sets, data mining is the computational process for discovering new patterns. Data mining provides major advantage in agriculture for disease detection, problem prediction and for optimizing inputs like pesticides, irrigation, fertilizers, etc. With advancement in technological applications in agriculture, a lot of information is made available. Hence, data mining techniques in agriculture is used for pattern reorganization and crop health detection. Reliable and timely estimates of crop production are important for taking various decisions for marketing, pricing, storage, distribution and import-export. The crop yields primarily depend on diseases, pests, climatic conditions, time of harvest, etc. As such, these predictions are very useful for agriculture domains. Data mining techniques are used not only for quantifying requirements of inputs and timely executing various agricultural operations but also for pre-harvest forecasting for crop yields. Data mining is also called as knowledge discovery database (KDD).

Data mining tasks can be classified into two categories:

- Descriptive data mining.
- Predictive data mining.

Descriptive data mining tasks characterize the general properties of the data in the database while predictive data mining is used to predict the direct values based on patterns

determined from known results. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest. As far as data mining technique is concern, in most of the cases, predictive data mining approach is used. Predictive data mining technique is used to predict future crop, weather forecasting, pesticides and fertilizers to be used, revenue to be generated and so on. [4].

### III. SURVEY OF LITERATURE

Many research studies have focused on the importance of data mining to be used as a tool for analysis of big data of agriculture for meaningful conclusions [1], [2].

Ashok Kumar and N. Kannathasan dealt with various data mining procedures that can be utilized in farming. Their examination suggested that a correlation of various data mining strategies could deliver a productive calculation for soil grouping [3].

M. C. Geetha researched Data Mining Techniques in Agriculture and argued that information mining in agriculture is quite useful to forecast the productivity and production of crops. She talked about various information-digging applications for tackling distinctive farming related issues [8].

Ruß G. utilized information got from three fields of Germany. Researcher utilized regression methods on farming yield information and reasoned that help vector regression can fill in as a superior reference demonstrates for yield expectation. Likewise, the model parameters which have been built upon one informational index can be utilized for different techniques on selected agriculture data [12].

S. Veenadhari, B. Misra, and C. Singh endeavored to aggregate the examination discoveries of various scientists who took a shot at harvest profitability information. The machine learning approach of coordinating software engineering with farming will help in gauge farm yields adequately [13].

A study by Ekasingh B. S. Ngamsomsuke K. Letcher R. A. & Spate J.M. has analyzed how data mining may be applied for the purposes of crop production [7]. Majority of earlier researchers including Dunstan D. (2009) who used data mining as a supportive tool with statistical analysis, have focused on crop yield management and its' quality evaluation [6].

Raorane A.A. and Kulkarni R.V. (2012) talked about different data mining strategies, as a result of use of data mining methods; an effective production system can be derived that can take care of complex farming issues [11].

Ramesh Vamanan and K. Ramar (2011) presumed that the Data Mining method (Naïve Bayes Classifier) when connected to a farming soil profile may enhance the confirmation of legitimate soil profile, substantial examples and profile classification are contrasted with standard statistical investigation strategies [10].

### IV. MATERIALS AND METHODOLOGY

#### A. Data Acquisition

The focal point of this research was to look at the effect of temperature, rainfall and soil parameters (namely soil type, Soil

pH, Carbon, phosphorus, and potassium) on the cotton yields for different farming locales in the study area.

For the present research, 27 talukas of the three districts of Gujarat State were taken. The soil parameter data were retrieved from the Soil Health Card data of Government of Gujarat routed through Anand Agricultural University, District Anand, Gujarat State. The production data were collected from the Department of Agriculture, Government of Gujarat, by approaching them personally and also by random access to farmers for verification of the information. The data of rainfall and temperature were collected from the Sardarkrushinagar Dantiwada Agricultural University, SKNagar, District Banaskantha, Gujarat State.

In the present study, the rainfall, temperature and five soil parameters were considered as independent factors and their effect on cotton productivity was analyzed. As a consequence, instances of cotton yield were examined against these datasets. The average rainfall, temperature and selected soil parameters of the ten years for each of the 21 talukas in the three districts were obtained from a secondary database. These data sets of ten years (from 2006 to 2015) were analyzed. The dataset was structured and combined to be managed in excel spreadsheet as talukas, average rainfall, maximum and minimum temperature, soil parameters, and cotton yield.

#### B. Analytical Procedure

To conduct different research experiments and calculations WEKA, Revolution R and SPSS have been used. Initially the data was collected and maintained in Microsoft Excel. Further, data transformation and other calculations were done in software like SPSS and Microsoft Excel. The total datasets was classified and kept in different folders; and further that ordered into the rainfall, temperature, and soil variation (pH, Carbon, Phosphorous, and Potassium) and yield. Different algorithms were tested in WEKA software to check and decide the most suitable among all algorithms and evaluate output with other datasets. These data were used in WEKA and R software for dept. analysis and experiments.

#### C. The Activity Experiments

The following examinations were done iteratively for calculating the impact of the climatic factors on cotton yield:

- 1) Rainfall and soil type relationship in view of the cotton crop.
- 2) Effect of nutritional variations of soil on cotton yield.

Further the data were restructured from the perspective of having sufficient depth and substance to be believable before initiating the processing. The lateral development happened because attributes were included from other data collection. Both the reduction and extension add up to the pre-preparing of the dataset and incorporation and delineation of outliers through Exploratory Visual Data Mining (EVDM) of the mapped information.

The compiled data were administered to techniques of combination, categorization, accumulation and statistical projection to locate elite method and related best fit design. In the auxiliary investigation through OLAP appraised for

association in the data sets. Examination of these outcomes were taken up and concluded accordingly.

At last, the two assortments of outcomes were combined and arranged diagrammatically to decide a general example that gave both a minute and apparent context of data. The datasets were then examined through a progression of tests which included cross classifications, correlation, sequencing, time series analysis and regression. The outcomes of these tests were then investigated and delineated in detail.

As a result of this study, recommendations for getting good cotton crop yields by managing the agronomical and other practices in lieu of the difficult weather and soil conditions were made for north Gujarat farming area.

#### D. Data Analysis

The information utilized in this investigation had diverse attributions and was comprised of five separate however related substances. The majority of the datasets were in connection with North Gujarat, India.

In the current research, different data were used with varied reasoning. As referred in Section 3.3, the different datasets comprised of temperature, soil type, rainfall and soil parameters were used. Different data sets were obtained either from secondary source of data namely Department of Agriculture, Government of Gujarat; Anand Agricultural University, Anand and Sardarkrushinagar Dantiwada Agricultural University, Sardarkrushinagar, Gujarat, India. The counter verification of these data and obtaining few other data were done by personal deliberation with scientists and farmers. All the datasets were fitted explicitly for the cotton crop production areas of the selected locale of the study.

The data extraction and pertinence involved using a number of software tools that represented an admixture of retrieving, pre-processing, scrutiny, data mining and revelation of temperature, soil type, rainfall, and soil parameters' data. The process was alienated into five fundamental stages namely data collection, pre-processing, handling, data examination and processing.

Different crop have their critical and optimum climatic requirements, for example, the increasing temperature may affect agriculture by reducing the productivity on different crops in different seasons. The analysis of climatology at the region level is most useful for the solution of practical agricultural problems. Temperature has a complex relationship to the development of plant at different growth stage.

#### E. Basic Model Process Flow

The entire information extraction and analysis process is demonstrated graphically in Fig. 1 below. It framed the initial segment of research in anticipation of the data collection, data preparation, data modeling and data storage in the protected investigations.

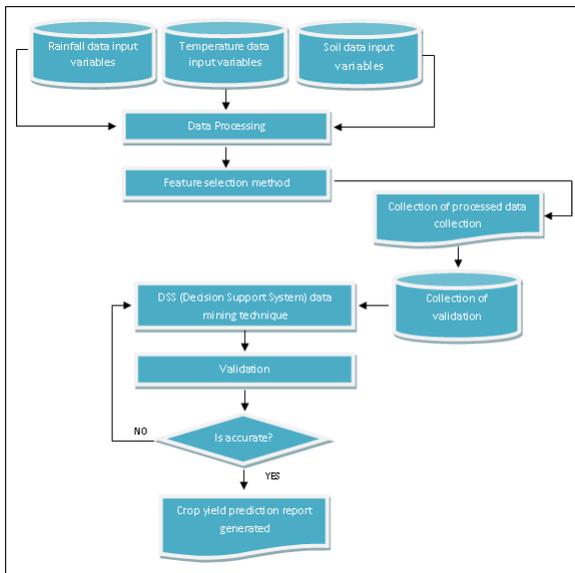


Fig. 1. Basic Model of Process Flow.

### F. Analytical Procedure

To conduct different research experiments and calculations WEKA, Revolution R and SPSS have been used. Initially the data was collected and maintained in Microsoft Excel. Further, data transformation and other calculations were done in software like SPSS and Microsoft Excel. The total datasets was classified and kept in different folders; and further that ordered into the rainfall, temperature, and soil variation (pH, Carbon, Phosphorous, and Potassium) and yield. Different algorithms were tested in WEKA software to check and decide the most suitable among all algorithms and evaluate output with other datasets. These data were used in WEKA and R software for dept. analysis and experiments.

- The Activity Experiments

The following examinations were done iteratively for calculating the impact of the climatic factors on cotton yield:

- 1) Rainfall and soil type relationship in view of the cotton crop.
- 2) Impact of rainfall on cotton yield.
- 3) Impact of temperature on cotton yield.
- 4) Combined effect of rainfall and temperature on cotton yield.
- 5) Effect of nutritional variations of soil on cotton yield.

As a result of this study, recommendations for getting good cotton crop yields by managing the agronomical and other practices in lieu of the difficult weather and soil conditions were made for north Gujarat farming area.

## V. RESULTS AND DISCUSSION

### A. Taluka Wise Yield Prediction for Low and High Rainfall Years

The data sets of low and high rainfall years (2009 and 2013 respectively) were applied for analytical process to WEKA. The datasets used and calculations made in WEKA are given in Table I.

TABLE I. WEKA PREDICTION FOR YEAR 2009 AND 2013

| Taluka      | LR 2009         |              | HR 2013         |              |
|-------------|-----------------|--------------|-----------------|--------------|
|             | Predicted Yield | Actual Yield | Predicted Yield | Actual Yield |
| AMIRGHADH   | 649.52          | 488          | 627.46          | 598          |
| BECHARAJI   | 481.23          | 584          | 620.87          | 568          |
| BHABHAR     | 500.49          | 526          | 564.92          | 523          |
| DANTA       | 401.41          | 456          | 392.28          | 395          |
| DANTIWADA   | 440.27          | 556          | 610.46          | 630          |
| DEESA       | 657.96          | 621          | 615.55          | 630          |
| DHANERA     | 599.29          | 578          | 647.30          | 615          |
| DIYODAR     | 594.51          | 543          | 645.93          | 599          |
| HIMATNAGAR  | 386.82          | 599          | 605.44          | 617          |
| IDAR        | 737.05          | 636          | 557.37          | 601          |
| KADI        | 427.99          | 500          | 613.31          | 594          |
| KANKARAJ    | 514.21          | 513          | 619.88          | 597          |
| KHEDBHRAMHA | 373.41          | 408          | 435.85          | 451          |
| KHERALU     | 211.74          | 300          | 370.83          | 350          |
| MAHESANA    | 579.97          | 548          | 645.34          | 602          |
| PALANPUR    | 604.30          | 623          | 627.25          | 645          |
| POSHINA     | 475.79          | 348          | 375.57          | 303          |
| PRANTIJ     | 508.20          | 615          | 573.36          | 536          |
| SATALASANA  | 289.91          | 296          | 369.98          | 366          |
| TALOD       | 535.21          | 642          | 537.68          | 566          |
| UNJHA       | 365.22          | 311          | 387.25          | 369          |
| VADALI      | 494.03          | 598          | 633.98          | 565          |
| VADGAAM     | 384.59          | 508          | 352.14          | 439          |
| VADNAGAR    | 601.71          | 578          | 576.96          | 754          |
| VIJAPUR     | 543.93          | 721          | 721.76          | 685          |
| VIJAYNAGAR  | 448.53          | 568          | 539.17          | 495          |
| VISNAGAR    | 597.04          | 654          | 632.21          | 684          |

The difference between actual and predicted yield for the low rainfall year (2009) in different talukas varied greatly. It is obvious that the maximum difference in predicted and actual yield was observed for Poshina taluka as 36.72 per cent, followed by Amirghadh taluka (33.10 per cent) and Unjha taluka (17.43 per cent). The average difference between maximum value of predicted yield and actual yield was 36.72 per cent. The average difference between minimum value of predicted yield and actual yield was -35.42 per cent. However, the difference between average of predicted and actual yields of all talukas was only -5.39 per cent.

The difference between actual and predicted yield for the high rainfall year (2013) in different talukas varied greatly. It is obvious that the maximum difference in predicted and actual yield was observed for Poshina taluka as 23.95 per cent, followed by Vadali taluka (12.21 per cent) and Becharaji taluka (9.31 per cent). The average difference between maximum value of predicted yield and actual yield was 23.95 per cent. The average difference between minimum value of predicted yield and actual yield was -23.48 per cent. However, the difference between average of predicted and actual yields of all talukas was only 1.55 per cent.

### B. Taluka Wise Yield Prediction for Low and High Temperature Years

The difference between actual and predicted yield for the low temperature year (2015) in different talukas varied greatly.

It is obvious that the maximum difference in predicted and actual yield was observed for Becharaji taluka as 18.03 per cent, followed by Vadnagar taluka (17.01 per cent) and Satlasana taluka (16.27 per cent). The average difference between maximum value of predicted yield and actual yield was 18.03 per cent. The average difference between minimum value of predicted yield and actual yield was -29.24 per cent. However, the difference between average of predicted and actual yields of all talukas was only 0.44 per cent.

The difference between actual and predicted yield for the high temperature year (2010) [5] in different talukas varied greatly. It is obvious that the maximum difference in predicted and actual yield was observed for Talod taluka as 34.88 per cent, followed by Prantij taluka (26.02 per cent) and Vijaynagar taluka (25.05 per cent). The average difference between maximum value of predicted yield and actual yield was 34.88 per cent. The average difference between minimum value of predicted yield and actual yield was -38.34 per cent. However, the difference between average of predicted and actual yields of all talukas was only -1.87 per cent.

The data sets of high and low temperature years (2010 and 2015 respectively) were applied for analytical process to WEKA. The datasets used and calculations made in WEKA are given in Table II.

TABLE II. WEKA PREDICTION FOR YEAR 2010 AND 2015

| Taluka      | HT 2010         |              | LT 2015         |              |
|-------------|-----------------|--------------|-----------------|--------------|
|             | Predicted Yield | Actual Yield | Predicted Yield | Actual Yield |
| AMIRGHADH   | 502.43          | 564          | 629.09          | 752          |
| BECHARAJI   | 558.29          | 528          | 634.99          | 538          |
| BHABHAR     | 490.08          | 547          | 544.97          | 518          |
| DANTA       | 388.33          | 383          | 394.86          | 558          |
| DANTIWADA   | 578.35          | 485          | 665.49          | 621          |
| DEESA       | 631.26          | 606          | 613.14          | 589          |
| DHANERA     | 555.71          | 560          | 640.64          | 604          |
| DIYODAR     | 534.43          | 630          | 595.15          | 592          |
| HIMATNAGAR  | 659.87          | 575          | 638.67          | 623          |
| IDAR        | 404.08          | 513          | 608.32          | 627          |
| KADI        | 498.56          | 537          | 560.68          | 540          |
| KANKARAJ    | 394.64          | 640          | 590.18          | 640          |
| KHEDBHRAMHA | 464.23          | 451          | 440.14          | 434          |
| KHERALU     | 278.17          | 320          | 361.47          | 330          |
| MAHESANA    | 539.36          | 574          | 616.16          | 603          |
| PALANPUR    | 520.94          | 630          | 640.54          | 733          |
| POSHINA     | 345.32          | 388          | 378.34          | 368          |
| PRANTIJ     | 626.34          | 497          | 594.32          | 652          |
| SATALASANA  | 303.52          | 315          | 377.89          | 325          |
| TALOD       | 698.69          | 518          | 587.47          | 620          |
| UNJHA       | 300.64          | 334          | 378.01          | 350          |
| VADALI      | 594.92          | 526          | 618.91          | 629          |
| VADGAAM     | 278.92          | 364          | 486.74          | 540          |
| VADNAGAR    | 583.01          | 670          | 772.29          | 660          |
| VIJAPUR     | 753.58          | 694          | 700.61          | 666          |
| VIJAYNAGAR  | 573.98          | 459          | 542.80          | 589          |
| VISNAGAR    | 640.13          | 674          | 702.45          | 673          |

## VI. DISCUSSION

Prediction of yield of a crop is very difficult; as it is the sum of complex interrelationship of many factors. The water affects a lot on cotton crop production. Though the farmers have no control on precipitation; although if having facilities of irrigation; timely irrigation, method of irrigation, quantity of irrigation and other irrigation management issues plays very important role in cotton crop production. Even delaying irrigation by one or two days in a peak season alters the effect of insect and pest infestation on the crop. In the current research, the data mining classification function of Gaussian Processes showed strong positive correlation between the average annual rainfall and cotton crop yield for the selected 27 talukas.

If real time outputs of this information are communicated to farmers for improving their crop management, it can really contribute to sustainable as well as improved production.

## VII. CONCLUSION

Data mining is the process of finding the useful outcomes from the large data sets. During this work, the time series forecasting package have been used for regression approach of Gaussian Processes for yield prediction. The Gaussian Processes algorithm using with parameters like year, crop yield, temperature, rainfall and the five soil parameters are considered within the model development.

Another important factor for cotton crop production is temperature. Temperature in air and soil; difference in day and night temperature; sudden changes in temperature; etc affects the productivity. The hot winds, at a particular growth period are harmful for and at a different growth period is useful for production. When temperatures become too hot, fertilization may be compromised, leading to fewer seeds produced per boll, smaller boll masses, and ultimately, lint yield reductions [15]. Temperature affects cotton crop in a complex way on the production of the crop. However, the results of this study shows very strong and positive correlation between air temperature (minimum and maximum) on cotton crop yield.

The soil parameters, variety, farmers' management abilities and their socio-economic capabilities, etc all independently as well as their interactive effects decides the production of cotton. The yield was directly associated with improved soil water relations resulting from the cropping and tillage treatments. Application of varying levels of fertilizer in combination with bio-fertilizer positively influences to cotton yield as they improve the availability of NPK to the crop. The soil parameters including soil pH, SOC, P & K have positive correlation with cotton crop production. As the soil type in the entire operational area was identical (sandy loam type – Goradu type), its' correlation with cotton crop yields could not be assessed.

If real time outputs of this information are communicated to farmers for improving their crop management, it can really contribute to sustainable as well as improved production.

## VIII. WAY FORWARD

Forecasts of production include imbedded assumptions about farmers' reactions to changes in output prices, input

prices, weather forecasts, labor availability, input availability, storage capacity, marketing opportunities, food security, and changes in technology, government policies and other innumerable factors. Forecasts of consumption are really forecasts of textile mill managers' choices in response to the welter of price information, resource constraints and government policies they face, etc.

Nevertheless, even though human behavior is highly variable and unanticipated policy shocks are common, with great advances in technology improvement in forecasts has not been. Perhaps this is because the information we are getting faster is actually degrading in quality. In other words, the statistics on which forecasts are based are becoming less accurate, thus undermining the value of getting those statistics more easily.

#### REFERENCES

- [1] Abdullah, A., and Ansari, I. A. (2005), Discovery of cropping regions due to Global Climatic Changes using Data Mining, Paper presented at the 3rd International Symposium on Intelligent Information Technology in Agriculture, Beijing.
- [2] Chien, and Chena (2008), Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry Expert Systems with Applications, 34(1, January 2008), 280-290.
- [3] D, Ashok Kumar and Kannathasan, N (2011) A Survey on Data Mining and Pattern Recognition Techniques for Soil Data Mining, International Journal of Computer Science Issues, Vol. (8).
- [4] Developing innovative applications in agriculture using data mining-Sally Jo Cunningham and Geoffrey Holmes.
- [5] Drew, J. (2010), Operating In a Change Environment – NEAR/Drought Reform.
- [6] Dunstan, D. (2009), Hierarchies of sustainability in a catchment, Paper presented at the 4th International Conference on Sustainable Development and Planning, Cyprus.
- [7] Ekasingh, B. S. et al (2005), A Data Mining approach to simulating land use decisions: Modelling farmer's crop choice from farm level data for integrated water resource management, Paper presented at the Proceedings of the 2005 International Conference on Simulation and Modelling.
- [8] Geetha, M.C. (2018), A Survey and Analysis on Regression Data Mining Techniques in Agriculture, International Journal of Pure and Applied Mathematics, Vol 118 No. 8, 341-347.
- [9] Gleaso CP. Large area yield estimation/forecasting using plant process models. Paper presentation at the winter meeting American society of agricultural engineers palmer house, Chicago, Illinois. 1982; 14–17.
- [10] Ramesh Vamanan, K. Ramar (2011), Classification of agricultural land soils a data mining approach, International Journal on Computer Science and Engineering, Vol. 3(1).
- [11] Raorane A.A., Kulkarni R.V. (2012), Data Mining: An effective tool for yield estimation in the agricultural sector, International Journal of Emerging Trends and Technology in Computer Science (IJETTCS), Vol (1-2).
- [12] Ruß, G. et al (2009), Visualization of agriculture data using self-organizing maps, Paper presented at the Proceedings of AI-2008.
- [13] S. Veenadhari, B. Misra and C. Singh (2014), "Machine learning approach for forecasting crop yield based on climatic parameters," 2014 International Conference on Computer Communication and Informatics, Coimbatore, 2014, pp. 1-5, doi: 10.1109/ICCCI.2014.6921718.
- [14] Veenadhari S, Misra B, Singh CD. Data mining techniques for predicting crop productivity—A review article. In: IJCST. 2011; 2(1).
- [15] William T. Pettigrew (2008), Potassium influences on yield and quality production for maize, wheat, soybean and cotton, Physiologia Plantarum, Volume 133, Issue 4.

#### ANNEXURE I

The following are some of the more frequent types of information that can be derived from the basic data:

##### A. Air temperature

- 1) Temperature probabilities;
- 2) Chilling hours;
- 3) Degree-days;
- 4) Hours or days above or below selected temperatures;
- 5) Interdiurnal variability;
- 6) Maximum and minimum temperature statistics;
- 7) Growing season statistics, that is, dates when threshold temperature values for the growth of various kinds of crops begin and end.

##### B. Precipitation

- 1) Probability of a specified amount during a period;
- 2) Number of days with specified amounts of precipitation;
- 3) Probabilities of thundershowers;
- 4) Duration and amount of snow cover;
- 5) Dates on which snow cover begins and ends;
- 6) Probability of extreme precipitation amounts.

##### C. Wind

- 1) Windrose;
- 2) Maximum wind, average wind speed;
- 3) Diurnal variation;
- 4) Hours of wind less than selected speed.

##### D. Sky cover, sunshine, radiation

- 1) Per cent possible sunshine;
- 2) Number of clear, partly cloudy, cloudy days;
- 3) Amounts of global and net radiation.

##### E. Humidity

- 1) Probability of a specified relative humidity;
- 2) Duration of a specified threshold of humidity.

##### F. Free water evaporation

- 1) Total amount;
- 2) Diurnal variation of evaporation;
- 3) Relative dryness of air;
- 4) Evapotranspiration.

##### G. Dew

- 1) Duration and amount of dew;
- 2) Diurnal variation of dew;
- 3) Association of dew with vegetative wetting;
- 4) Probability of dew formation based on the season.

##### H. Soil temperature

- 1) Mean and standard deviation at standard depth;
- 2) Depth of frost penetration;
- 3) Probability of occurrence of specified temperatures at standard depths;
- 4) Dates when threshold values of temperature (germination, vegetation) are reached.

*I. Weather hazards or extreme events*

- 1) Frost;
- 2) Cold wave;
- 3) Hail;
- 4) Heatwave;
- 5) Drought;
- 6) Cyclones;
- 7) Flood;
- 8) Rare sunshine;
- 9) Waterlogging.

*J. Agrometeorological observations*

- 1) Soil moisture at regular depths;
- 2) Plant growth observations;
- 3) Plant population;
- 4) Phenological events;
- 5) Leaf area index;
- 6) Above-ground biomass;
- 7) Crop canopy temperature;
- 8) Leaf temperature;
- 9) Crop root length.

# Data Analysis of Coronavirus CoVID-19: Study of Spread and Vaccination in European Countries

Hela Turki, Kais Khrouf  
Jouf University, Sakaka  
Saudi Arabia

**Abstract**—Humanity has gone since a long time through several pandemics, such as: H1N1 in 2009 and also Spanish flu in 1917. In December 2019, the health authorities of China detected unexplained cases of pneumonia. The WHO World Health Organization has declared the apparition of CoVID-19 (novel Coronavirus) that caused a global pandemic in 2020. In data analysis, multiple approaches and diverse techniques were used to extract useful information from multiple heterogeneous sources and to discover knowledge and new information for decision-making; it is used in different business and science domains. In this context, we propose to use the multidimensional analysis techniques based on two concepts: fact (subject of analysis) and dimensions (axes of analyses). This technique allows decision makers to observe data from various heterogeneous sources and analyze them according several viewpoints or perspectives. More precisely, we propose a multidimensional model for analyzing the Coronavirus CoVID-19 data (spread and vaccination in European countries). This model is based on constellation schema that contains several facts surrounded by common dimensions.

**Keywords**—Multidimensional model; constellation schema; coronavirus covid-19; vaccination; European countries

## I. INTRODUCTION

Since December 2019, the new cases of pneumonia were detected in Wuhan City (Hubei Province of China). This novel virus caused the new infectious respiratory disease, called Covid-19 by the World Health Organization (WHO) [19] (Pandemic in 2020 with millions of deaths around the world). The fight against this global pandemic is causing cancellations of sporting and cultural events, the implementation of containment measures and the closure of the borders by many countries, etc. It also has effects in terms of social and economic instability [14].

In order to slow the contagion of this new virus, several studies were proposed in the literature [12][15][16][17], especially about the spread of the Coronavirus [13]; statistics are announced every day by the countries and databases have been established to store this data. In this paper, we propose to use the Multidimensional Analysis techniques in order to analyze the spread of Coronavirus Covid-19 and the evolution of vaccination in European Countries. This technique allows decision makers to observe data from various sources and analyze them according to several viewpoints. A multidimensional model is composed into two concepts: Dimension and Fact. Dimensions contain a set of unique values in order to categorize a particular theme (Countries, Dates,

etc.). Fact is a subject of analysis and it is described by a set of measures.

This paper presents a new approach based on the use of multidimensional techniques on Coronavirus Covid-19 data and the user-defined constraints based on colors in order to highlight relevant information.

This paper is organized as follows. Section 2 presents the literature review for spreading of Coronavirus Covid-19 (Works about data analysis). Then, we present the phase of data preparation (Extraction, Cleaning, Transformation and Loading of Data). In Section 4, we propose a data warehouse schema for storing the prepared data. The next section describes the multidimensional model we propose for analyzing the spread and the vaccination of Coronavirus Covid-19 data. Finally, we present the phase of implementation for European countries and then Conclusion.

## II. LITERATURE REVIEW

Since the appearance of the Coronavirus Covid-19, several studies have focused on the spread of the virus (Medical [13] or Data Analysis aspects [18]).

The objective of [1] is to examine the correlation between pollution and climate data and the Covid-19 pandemic. They propose a data warehouse and data cubes built on data from the regions of Lombardy and Puglia (Italia). Their results show that the Covid-19 pandemic is spreading significantly in regions characterized by the absence of rain and wind.

In [2], the authors study the relationship between new cases of Coronavirus Covid-19 and the Multidimensional Poverty Index (MPI) in the city of Manizales (Colombia). The results of the exploration indicate that in the communes of greater poverty the density of cases per Covid-19 is greater; the relation exists between these two parameters.

Internet of Things (IoT) is an interconnection of Internet and physical devices. These devices are record, monitor and respond. The use of IoT with smart sensors to measure and record the body temperature of individuals can help to identify the infected and to maintain social distance. The authors of [3] propose an IoT architecture in order to minimize the spreading of Covid-19.

In [4], the authors study the evolution of cases and deaths of Covid-19 compared to the population of Brazilian cities. The results show that in the short term small towns are proportionately more affected by Covid-19 during the initial

spread of the disease. In the long term, large cities begin to have a higher incidence of cases and deaths.

The authors of [5] propose an interactive visualization using the concept of Tableau [6] for analyzing data of Covid-19. A Tableau is used to show the personalized and the most important data (dashboards and worksheets). They consider that data analysis can be very fast with Tableau and Visualizations (several visualizations in a single view).

The author of [7] presents a data analysis of Covid-19 in cities of China, by using datasets. He uses a correlation matrix for the phase of data preparation (to summarize data). He uses Python libraries Matplotlib and Seaborn for visualizing data.

In this paper, we propose a multidimensional model based on Constellation Model in order to study the spread of Coronavirus Covid-19 in European countries and the evolution of vaccination, according to several dimensions. The first stage concerns the data preparation (presented in next section).

### III. DATA PREPARATION

Data preparation is the process of several steps (gathering, combining and structuring data) in order to analyze them in business intelligence and data visualization applications. Fig.1 presents the process we propose for data preparation: Data extraction, Data cleaning, Data transformation and Data loading.

#### A. First Step: Data Collection

In this paper, the data used was extracted from [20]. The period of analysis is between 01/01/2021 and 30/09/2021. We mainly use the following files:

The first file concerns data on testing for Covid-19 by week and country and contains the following data: Country name and code, Week of year, Level (national or sub-national), Region code and name, Number of new confirmed cases, Population, Testing rate per 100000 population, Positivity rate and Source.

The second file concerns data on Covid-19 vaccination and contains the following data: Week of year, Country code, Population denominators for target groups, Number of doses received, Number of first dose vaccine, Number of individuals refusing the first vaccine dose, Number of second dose vaccine, Number of doses where the type of dose was not specified, Region, Target group, Name of vaccine and Population.

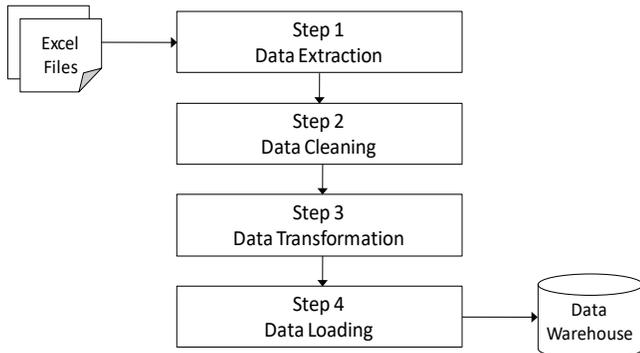


Fig. 1. Data Preparation.

#### B. Second Step: Data Cleaning

In this step, we removed unnecessary data:

- Source and Level from the first file.
- Population denominators for target groups and Target group from the second file.

We also add the following data in order to perform analyzes at several levels of granularity:

- Month, Trimester and Year for the week of year.
- Zone for countries: we distinguish four zones: Eastern Europe, Western Europe, Northern Europe and Southern Europe (cf. Table I).
- Continent: In this study, we focus on Europe.

TABLE I. ZONES OF EUROPE

| Zones           | Countries                                                                                                                   |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------|
| Eastern Europe  | Belarus, Bulgaria, Czech, Hungary<br>Poland, Moldova, Romania, Russia, Slovakia, Ukraine                                    |
| Western Europe  | Austria, Belgium, France, Germany<br>Liechtenstein, Luxembourg, Monaco<br>Netherlands, Switzerland.                         |
| Northern Europe | Denmark, Estonia, Finland, Iceland<br>Ireland, Latvia, Lithuania, Norway,<br>Sweden, United Kingdom.                        |
| Southern Europe | Albania, Andorra, Bosnia & Herzegovina, Croatia,<br>Greece, Italy, Malta, Portugal, Serbia, Slovenia<br>Spain and Macedonia |

#### C. Third Step: Data Transformation

In this phase, we merged the following data from the two files: Country code, Year of week, Region and Population.

The result after cleaning and merging data is a new file that contains:

- Week of year, Month, Trimester and Year.
- Country Name and Code, Population, Region code and Name.
- Number of new confirmed cases, Testing rate per 100000 population and Positivity rate.
- Number of doses received, Number of first dose vaccine, Number of individuals refusing the first vaccine dose, Number of second dose vaccine, Number of doses where the type of dose was not specified and Name of vaccine.

#### D. Four Step: Data Loading

After data is retrieved, extracted and transformed, it is then loaded into a storage system (a data warehouse); it involves sorting, checking integrity, and building indices and partition.

After the initial load, the data warehouse needs to be updated by the incremental changes in the data sources.

#### IV. DATA WAREHOUSING

Data storage is keeping data in a secure location that the user can easily access. An operational database handles frequent daily changes due to the transactions that take place by the company. However, a data warehouse provides consolidated data in multidimensional form. [8].

A data warehouse is constructed by heterogeneous data from multiple sources in order to support analytical reporting and decision making [11]. Indeed, it focuses on modeling and analysis of data to help decision-makers. The data in the warehouse must be subject oriented, integrated and non-volatile.

The data warehouse possesses consolidated historical data in order to organize, use and analyze this data to take strategic decisions. The main objective of data warehouses is to transform heterogeneous data into a form suitable for analysis.

In this step, we propose a schema of data warehouse (cf. Fig. 2) by using the class diagram of UML (Unified Modeling Language) that contains six classes: Date, Zone, Country, Region, Testing and Vaccine.

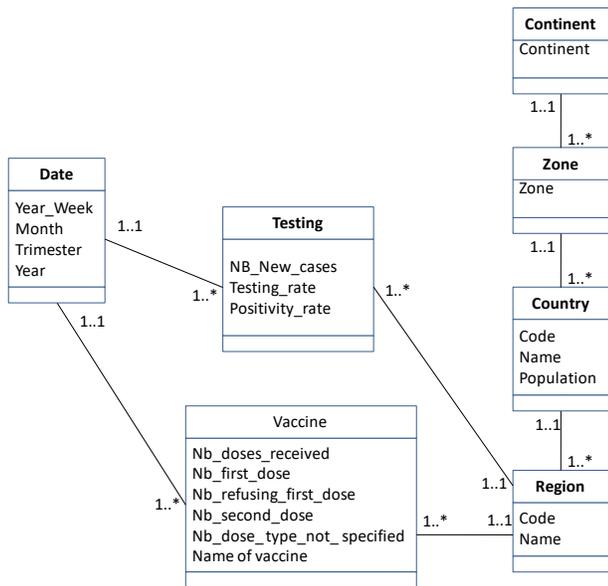


Fig. 2. Data Warehouse for Testing and Vaccination.

#### V. DATA ANALYSIS: MULTIDIMENSIONAL SCHEMA

A data warehouse provides Online Analytical Processing (OLAP) tools that present an interactive analysis of data in a multidimensional view. The results of these OLAP tools are generally Data Cubes, defined by dimensions (described by attributes and hierarchies) and facts (described by measures) [9].

- A dimension is a structure that describes a subject in order to help decision-makers answer business questions (Example: product, store, and date).
- An attribute describes a summary level or characteristic of a dimension (Example: Year).

- A hierarchy classifies a dimension into several levels of granularity (Example: Date can be decomposed into Date→Month→Year).
- A fact presents a subject that models a set of events (Examples: sales, purchases); it has dynamic properties (numeric attributes).
- A measure is a numerical property of quantitative aspect that is relevant to analysis (Example: quantity, number\_of\_customers).

Schema represents a logical description of a database, data warehouse, XML document [10], etc. If a database generally uses relational model, a data warehouse can use Star, Snowflake or Constellation schemas.

- Star Schema: Each dimension is represented by only one-dimension table and the fact table at the center that contains the keys of all dimensions.
- Snowflake Schema: Some dimension tables are normalized.
- Fact Constellation Schema: It contains multiple fact tables connected by common dimensions. Table II presents the components of multidimensional schema we propose.

TABLE II. COMPONENTS OF MULTIDIMENSIONAL SCHEMA

| Concept                                                                     | Description                                                                                                        |
|-----------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| <b>Constellation C</b><br>$C = (F ; D_i)$                                   | $F$ is a set of facts.<br>$D_i$ is a set of dimensions.                                                            |
| <b>Fact F</b><br>$F = (NameFct ; M_i)$                                      | $NameFct$ is the fact name of $F$ .<br>$M_i$ is a list of measures.                                                |
| <b>Dimension <math>D_i</math></b><br>$D_i = (NameDim_i ; Att_j ; Hierar_k)$ | $NameDim_i$ is the dimension name.<br>$Att_j$ is the list of attributes.<br>$Hierar_k$ is the list of hierarchies. |

Fig. 3 presents the proposed multidimensional model that contains two facts (Testing and Vaccine) surrounded by two dimensions (D\_Date, D\_Region). D\_Date is decomposed into the hierarchy H1 (Week→Month→Trimester→Year). D\_Region is decomposed into the hierarchy H2 (Region→Country→Zone→Continent).

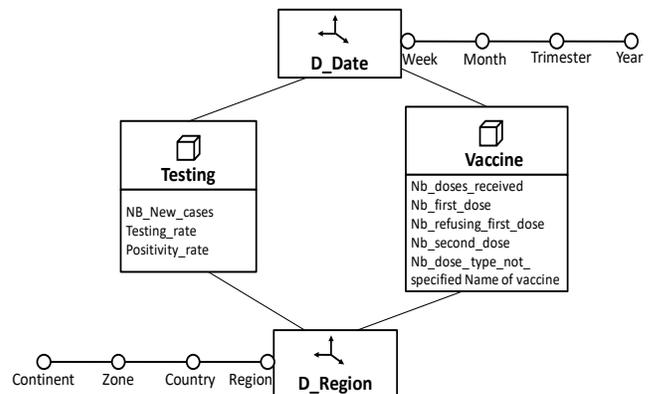


Fig. 3. Multidimensional Schema.

## VI. IMPLEMENTATION AND DISCUSSIONS

The data used for analysis in this paper was extracted from [20]. The analysis period is between 01/01/2021 and 30/09/2021. In this section, we present several examples of multidimensional queries for analyzing data about spreading of Coronavirus Covid-19 and vaccination.

The first query concerns the average of positivity rate by Zone and Trimester. Table III presents the result of this query.

Example 1:

|                                 |
|---------------------------------|
| Dimension 1: Zone               |
| Dimension 2: Trimester          |
| Fact: Average (Positivity_rate) |

TABLE III. AVERAGE OF POSITIVITY RATE BY ZONE AND TRIMESTER

|                 | T1-2021 | T2-2021 | T3-2021 |
|-----------------|---------|---------|---------|
| Eastern Europe  | 18,29   | 9,43    | 0,71    |
| Northern Europe | 3,34    | 2,00    | 1,81    |
| Southern Europe | 7,73    | 3,87    | 4,19    |
| Western Europe  | 5,29    | 3,78    | 2,79    |

We note that the positivity rate is very high for the zone of Eastern Europe during the first two trimesters of 2021. In order to analyze more this multidimensional table, we propose to apply the Drill-down Operator that fragments data into smaller parts. It can be done by descending from a level to another in the hierarchy. Example: For Dimension 1, we can visualize data by passing from Zone to Country (cf. Table IV).

Example 2:

|                                 |
|---------------------------------|
| Dimension 1: Country            |
| Dimension 2: Trimester          |
| Fact: Average (Test_Positivity) |

TABLE IV. AVERAGE OF POSITIVITY RATE BY COUNTRY AND TRIMESTER

|             | T1-2021 | T2-2021 | T3-2021 |
|-------------|---------|---------|---------|
| Austria     | 2,22    | 0,40    | 0,18    |
| Belgium     | 5,74    | 5,65    | 3,27    |
| Bulgaria    | 13,01   | 8,41    | 3,04    |
| Croatia     | 10,34   | 11,83   | 1,98    |
| Cyprus      | 1,13    | 0,72    | 0,99    |
| Czechia     | 11,93   | 0,94    | 0,22    |
| Denmark     | 0,52    | 0,15    | 0,41    |
| Estonia     | 14,93   | 6,98    | 4,44    |
| Finland     | 2,66    | 1,28    | 2,46    |
| France      | 6,68    | 4,91    | 3,98    |
| Germany     | 6,96    | 6,38    | 3,88    |
| Greece      | 3,63    | 0,87    | 1,54    |
| Hungary     | 14,78   | 8,50    | 1,10    |
| Iceland     | 0,38    | 0,26    | 1,56    |
| Ireland     | 7,70    | 2,45    | 5,88    |
| Italy       | 6,90    | 3,54    | 2,16    |
| Latvia      | 6,61    | 3,45    | 0,93    |
| Lithuania   | 8,83    | 4,35    | 3,55    |
| Luxembourg  | 1,92    | 1,43    | 1,42    |
| Malta       | 6,06    | 1,19    | 2,94    |
| Netherlands | 9,46    | 8,13    | 4,07    |
| Norway      | 1,38    | 1,44    | 1,40    |
| Poland      | 18,82   | 10,86   | 0,52    |
| Portugal    | 6,79    | 1,11    | 2,68    |
| Romania     | 13,53   | 5,61    | 1,47    |
| Slovakia    | 46,43   | 8,03    | 0,70    |
| Slovenia    | 6,06    | 1,93    | 0,95    |
| Spain       | 8,81    | 4,96    | 9,34    |
| Sweden      | 11,76   | 8,25    | 3,60    |

For analyzing the number of cases compared to population by Country and Month for Western Europe, we propose the following query. In this example, We add constraints on the measure of the multidimensional table based on colors ( $\leq 2\%$ : Green;  $\leq 3\%$ : Orange;  $> 4\%$ : Red) in order to highlight the most important values (cf. Table V).

Example 3:

|                                                           |
|-----------------------------------------------------------|
| Dimension 1: Month                                        |
| Dimension 2: Country                                      |
| Dimension 3: Zone {Western Europe}                        |
| Fact: (Number_Cases/Population)*1000                      |
| { $\leq 2\%$ : Green; $\leq 3\%$ : Orange; $> 4\%$ : Red} |

TABLE V. NUMBER OF CASES BY MONTH AND COUNTRY FOR WESTERN EUROPE

| Western Europe |         |         |        |         |            |             |
|----------------|---------|---------|--------|---------|------------|-------------|
|                | Austria | Belgium | France | Germany | Luxembourg | Netherlands |
| janv-21        | 1,38    | 1,30    | 1,74   | 1,34    | 1,50       | 2,30        |
| févr-21        | 1,24    | 1,34    | 1,98   | 0,68    | 1,90       | 1,54        |
| mars-21        | 2,10    | 2,28    | 2,17   | 1,01    | 2,13       | 2,35        |
| avr-21         | 1,93    | 2,37    | 2,77   | 1,55    | 2,17       | 3,00        |
| mai-21         | 0,92    | 1,65    | 1,01   | 1,06    | 1,36       | 2,37        |
| juin-21        | 0,23    | 0,71    | 0,67   | 0,21    | 0,36       | 0,76        |
| juil-21        | 0,12    | 0,52    | 0,60   | 0,07    | 0,90       | 1,71        |
| août-21        | 0,43    | 1,16    | 3,19   | 0,23    | 0,73       | 1,49        |
| sept-21        | 1,13    | 1,52    | 2,32   | 0,78    | 0,81       | 0,99        |

In the next example, we apply the same query for the zone of Eastern Europe (cf. Table VI).

Example 4:

|                                                           |
|-----------------------------------------------------------|
| Dimension 1: Month                                        |
| Dimension 2: Country                                      |
| Dimension 3: Zone {Eastern Europe}                        |
| Fact: (Number_Cases/Population)*1000                      |
| { $\leq 2\%$ : Green; $\leq 3\%$ : Orange; $> 4\%$ : Red} |

TABLE VI. NUMBER OF CASES BY MONTH AND COUNTRY FOR EASTERN EUROPE

| Eastern Europe |          |        |         |         |        |         |          |
|----------------|----------|--------|---------|---------|--------|---------|----------|
|                | Bulgaria | Cyprus | Czechia | Hungary | Poland | Romania | Slovakia |
| janv-21        | 0,56     | 1,80   | 5,63    | 1,02    | 1,33   | 1,16    | 6,60     |
| févr-21        | 1,02     | 1,00   | 5,91    | 1,64    | 1,32   | 0,96    | 6,88     |
| mars-21        | 2,94     | 2,65   | 6,47    | 5,33    | 3,02   | 1,76    | 4,92     |
| avr-21         | 2,47     | 4,93   | 2,42    | 3,34    | 3,49   | 1,38    | 2,11     |
| mai-21         | 0,69     | 2,69   | 0,88    | 0,79    | 0,66   | 0,37    | 0,85     |
| juin-21        | 0,16     | 0,52   | 0,18    | 0,14    | 0,09   | 0,06    | 0,24     |
| juil-21        | 0,07     | 5,65   | 0,11    | 0,03    | 0,02   | 0,02    | 0,06     |
| août-21        | 0,37     | 5,15   | 0,12    | 0,04    | 0,03   | 0,08    | 0,09     |
| sept-21        | 1,37     | 1,94   | 0,16    | 0,13    | 0,06   | 0,45    | 0,38     |

We note that the number of cases is higher in zone Eastern Europe than Western Europe; mainly in Cyprus, Czechia Hungary and Slovakia.

We now analyze vaccines (Second doses) by Country and Trimester (cf. Table VII).

Example 5:

|                         |
|-------------------------|
| Dimension 1: Country    |
| Dimension 2: Trimester  |
| Fact: Sum (SecondDoses) |

TABLE VII. SUM OF SECOND DOSES BY COUNTRY AND TRIMESTER

|               | T1-2021  | T2-2021  | T3-2021  |
|---------------|----------|----------|----------|
| Austria       | 1304688  | 6051042  | 7707459  |
| Belgium       | 1452645  | 5784945  | 8654586  |
| Bulgaria      | 211609   | 1340262  | 648475   |
| Croatia       | 233581   | 1640893  | 1322132  |
| Cyprus        | 88440    | 536712   | 460645   |
| Czechia       | 1571130  | 6238333  | 7342890  |
| Denmark       | 900590   | 2678019  | 5398098  |
| Estonia       | 147075   | 627317   | 443309   |
| Finland       | 364736   | 3070298  | 9160820  |
| France        | 10687536 | 54558541 | 98587748 |
| Germany       | 3953903  | 20467926 | 24342182 |
| Greece        | 1784184  | 7348828  | 7387333  |
| Hungary       | 1583766  | 7260894  | 1932295  |
| Iceland       | 52869    | 215571   | 179425   |
| Ireland       | 651935   | 2441616  | 3749838  |
| Italy         | 9027705  | 34698931 | 64225845 |
| Latvia        | 57248    | 851813   | 457016   |
| Liechtenstein | 1735     | 7932     | 9515     |
| Lithuania     | 452004   | 2069468  | 1643559  |
| Luxembourg    | 43508    | 336621   | 319917   |
| Malta         | 116516   | 468899   | 167250   |
| Netherlands   | 675565   | 4084539  | 5326585  |
| Norway        | 617754   | 2538436  | 4019710  |
| Poland        | 6769621  | 25450557 | 19874078 |
| Portugal      | 1519613  | 6765390  | 11437756 |
| Romania       | 2239536  | 10406398 | 1291069  |
| Slovakia      | 520086   | 2204046  | 1678675  |
| Slovenia      | 214139   | 890110   | 544107   |
| Spain         | 6950539  | 21686332 | 36015214 |
| Sweden        | 1567226  | 6911820  | 10184471 |

We note that the number of vaccines is very low for the first trimester. To improve the visibility of this observation, we propose the following multidimensional query. Table VIII presents the result of this query.

Example 6:

Dimension 1: Country  
 Dimension 2: Trimester  
 Fact: Average(SecondDoses)  
 { <=20% : Red; <=40%: Orange; >=60%: Green }

TABLE VIII. SUM OF SECOND DOSES BY COUNTRY AND TRIMESTER: RESULT WITH COLORS

|               | T1-2021 | T2-2021 | T3-2021 |
|---------------|---------|---------|---------|
| Austria       | 8,66    | 40,17   | 51,17   |
| Belgium       | 9,14    | 36,40   | 54,46   |
| Bulgaria      | 9,62    | 60,91   | 29,47   |
| Croatia       | 7,31    | 51,33   | 41,36   |
| Cyprus        | 8,15    | 49,43   | 42,42   |
| Czechia       | 10,37   | 41,17   | 48,46   |
| Denmark       | 10,03   | 29,83   | 60,13   |
| Estonia       | 12,08   | 51,52   | 36,41   |
| Finland       | 2,90    | 24,38   | 72,73   |
| France        | 6,52    | 33,30   | 60,18   |
| Germany       | 8,11    | 41,97   | 49,92   |
| Greece        | 10,80   | 44,48   | 44,72   |
| Hungary       | 14,70   | 67,37   | 17,93   |
| Iceland       | 11,80   | 48,13   | 40,06   |
| Ireland       | 9,53    | 35,68   | 54,80   |
| Italy         | 8,36    | 32,14   | 59,49   |
| Latvia        | 4,19    | 62,35   | 33,45   |
| Liechtenstein | 9,04    | 41,35   | 49,60   |
| Lithuania     | 10,85   | 49,69   | 39,46   |
| Luxembourg    | 6,22    | 48,09   | 45,70   |
| Malta         | 15,48   | 62,30   | 22,22   |
| Netherlands   | 6,70    | 40,49   | 52,81   |
| Norway        | 8,61    | 35,37   | 56,02   |
| Poland        | 12,99   | 48,85   | 38,15   |
| Portugal      | 7,70    | 34,30   | 57,99   |
| Romania       | 16,07   | 74,67   | 9,26    |
| Slovakia      | 11,81   | 50,06   | 38,13   |
| Slovenia      | 12,99   | 54,00   | 33,01   |
| Spain         | 10,75   | 33,54   | 55,71   |
| Sweden        | 8,40    | 37,03   | 54,57   |

## VII. CONCLUSION

Multidimensional analysis techniques have been used to visualize data from several perspectives, in order to help decision-makers exploring data according to several granularities and so make appropriate decisions.

In this paper, we propose a data warehouse for storing data about spreading of Coronavirus Covid-19 and vaccination in European countries. We present a multidimensional model based on constellation schema in order to deduce new knowledge. The user can add constraints or criteria on multidimensional tables based on colors in order to highlight the most important values.

This paper presents a new approach based on the use of multidimensional techniques on Coronavirus Covid-19 data and the user-defined constraints based on colors in order to highlight relevant information.

For future work, we plan to study the impact of vaccination on the spread of the Coronavirus Covid-19 by integrating statistical tools into multidimensional tables.

## REFERENCES

- [1] G. Agapito, C. Zucco, M. Cannataro, "COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data", International Journal of Environmental Research and Public Health, Vol. 17, No 15, 2020.
- [2] V. Henao-Cespedes, Y. A. Garcés-Gómez, S. Ruggeri, T. M. Henao-Cespedes, "Relationship analysis between the spread of COVID-19 and the multidimensional poverty index in the city of Manizales, Colombia", The Egyptian Journal of Remote Sensing and Space Science, 2021.
- [3] K. Krishna, K. Narendra, S. Rachna, "Role of IoT to avoid spreading of COVID-19, International Journal of Intelligent Networks, p. 32-35, 2020.
- [4] H. V. Ribeiro, A. S. Sunahara, J. Sutton, M. Perc, Q. S. Hanley, "City size and the spreading of COVID-19 in Brazil", PLoS One, Vol. 15, No 9, 2020.
- [5] N. Akhtar, N. Tabassum, A. Perwej, Y. Perwej, "Data analytics and visualization using Tableau utilitarian for COVID-19", Global Journal of Engineering and Technology Advances, Vol. 3, No. 2, p. 28-50, 2020.
- [6] V. Manohar, G. Arpan, B. Björn, "Tableau: A High-Throughput and Predictable VM Scheduler for High-Density Workloads", EuroSys Conference, ACM, New York, USA, 2018.
- [7] S. Shashank Raj, "Exploratory Data Analysis on outbreak of Coronavirus", International Research Journal of Engineering and Technology, Vol. 7, No. 2, 2020.
- [8] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, B. Becker, J. Caserta, "Kimball's Data Warehouse Toolkit Classics", 2<sup>nd</sup> Edition, Wiley, 2014.
- [9] T. B. Sardinha, M. V. Pinto, "Multi-Dimensional Analysis: Research Methods and Current Issues", Bloomsbury Academic, 2019.
- [10] K. Khrouf, H. Turki, "A Warehousing Approach of Semi-Structured Data", International Journal of Advanced Research in Engineering and Technology, Vol. 11, No 11, p. 1636-1644, 2020.
- [11] K. Khrouf, H. Turki, "Generic Multidimensional Model of Complex Data: Design and Implementation", International Journal of Computer Science and Network Security, Vol. 21 No. 12, 2021.
- [12] S. Kushwaha, S. Bahl, A. Kumar Bagha, K. Singh Parmar, M. Javaid, A. Haleem, R. Pratap Singh, "Significant Applications of Machine Learning for COVID-19 Pandemic", Journal of Industrial Integration and Management, Vol. 05, No. 04, p. 453-479, 2020.
- [13] L. Li, Z. Yang, Z. Dang, C. Meng, J. Huang, H. Meng, D. Wang, G. Chen, J. Zhang, H. Peng, "Propagation analysis and prediction of the COVID-19", Infectious Disease Modelling, Vol. 5, p. 282-292, 2020.

- [14] S. Shakibaei, G. C. de Jong, P. Alpkökina, T. H. Rashid, "Impact of the COVID-19 pandemic on travel behavior in Istanbul: A panel data analysis", *Sustainable Cities and Society*, Vol. 65, February 2021.
- [15] J. Sheng, J. Amankwah-Amoah, Z. Khan and X. Wang, "COVID-19 Pandemic in the New Era of BigData Analytics: Methodological Innovations and Future Research Directions", *British Journal of Management*, Vol. 32, p. 1164–1183, 2021.
- [16] C. Shorten, T. M. Khoshgoftar, B. Furht, "Deep Learning applications for COVID-19", *Journal of Big Data*, Vol. 8, Article Number 18, 2021.
- [17] A. Adiga, D. Dubhashi, B. Lewis, M. Marathe, S. Venkatramanan, A. Vullikanti, "Mathematical Models for COVID-19 Pandemic: A Comparative Analysis", *Journal of the Indian Institute of Science*, Vol. 100, p. 793–807, 2020.
- [18] W. Bo, Z. Ahmad, A. R. A. Alanzi, A. Ibrahim Al-Omari, E. H. Hafez, S. F. Abdelwahab, "The current COVID-19 pandemic in China: An overview and corona data analysis", *Alexandria Engineering Journal*, Vol. 61, No 2, p. 1369-1381, 2022.
- [19] <https://www.who.int/health-topics/coronavirus>
- [20] <https://www.ecdc.europa.eu/en>

# Design of Low Cost Bio-impedance Measuring Instrument

Rajesh Birok, Rajiv Kapoor

Dept. of Electronics & Communication Engineering  
Delhi Technological University Delhi, India

**Abstract**—It is a well-established fact that the electrical bio-impedance of a part of the human body can provide valuable information regarding physiological parameters of the human body, if the signal is correctly detected and interpreted. Accordingly, an efficient low-cost bio-electrical impedance measuring instrument was developed, implemented, and tested in this study. Primarily, it is based upon the low-cost component-level approach so that it can be easily used by researchers and investigators in the specific domain. The measurement setup of instrument was tested on adult human subjects to obtain the impedance signal of the forearm which is under investigation in this case. However, depending on the illness or activity under examination, the instrument can be used on any other part of the body. The current injected by the instrument is within the safe limits and the gain of the biomedical instrumentation amplifier is highly reasonable. The technique is easy and user-friendly, and it does not necessitate any special training, therefore it can be effectively used to collect bio-impedance data and interpret the findings for medical diagnostics. Moreover, in this paper, several existing methods and associated approaches have been extensively explored, with in-depth coverage of their working principles, implementations, merits, and disadvantages, as well as focused on other technical aspects. Lastly, the paper also deliberates upon the present status, future challenges and scope of various other possible bio-impedance methods and techniques.

**Keywords**—Noninvasive; bio-electrical; Impedance; bio-impedance; bio-medical; instrumentation

## I. INTRODUCTION

Extensive research is going on in the field of Bio-Medical Instrumentation. Researchers and investigators in this field are striving hard to find out new ways and methods for diagnosis and measurement of health parameters for the welfare of the mankind. There are two types of techniques for measuring biomedical signals namely non-invasive and invasive techniques. Non-invasive techniques are more suitable than the invasive ones if sufficient accuracy can be achieved using them. Whether, it is animal and plant cells or tissues, these are always made up of three-dimensional arrangement of cells and tissues. Therefore, human body is a complex biological structure and system, which is also made up of billions of cells and tissues arranged in 3-D formation [1]. The biological cells and tissues of both animals and plants floats in ECF which is known as Extra-Cellular Fluids. This ECF comprises Intra-Cellular Fluids (ICF) and Cell Membranes (CM) which may be with or without cell wall. When biological cells and tissues are subjected to the external electrical stimulus they respond and produces a complex bio-electrical impedance or simply known as bio-impedance. This bio-impedance is highly frequencydependent [2], [3].

Accordingly, frequency response of bio-impedance of cells and tissues of humans is greatly affected by physiological and physiochemical composition and structure of these cells and tissues. Moreover, it also changes from person to person. As a result, learning about cell and tissue anatomy and physiology through biological cell and tissue bio-impedance analysis will be a valuable resource. Therefore, it has been found that studying complex bio-impedance of biological cell and tissues is a useful method for non-invasive physiological and pathological investigations. As we know that the bio-electrical impedance of a biological cells or tissues is dependent on the signal frequency, however, multifrequency application may also be used for non-invasive diagnostics and medical investigations, so as to determine their physiological or pathological behavior or even properties. There are numerous Non-invasive bio-impedance techniques such as BIA (Bio-Impedance Analysis), EIT (Electrical Impedance Tomography), IPG (Impedance Plethysmography), ICG (Impedance Cardiography), etc. The bio-impedance measurement technique proposed in this research paper, is a low-cost, efficient, and effective non-invasive diagnostic technique.

## II. LITERATURE SURVEY

Impedance offered by a living tissue is known as bio-impedance. The broad variability of Cole parameters makes it difficult to use bio-impedance to distinguish animal and plant tissues. [4] defines a novel electronic procedure for distinguishing fruit or vegetable tissue. This system uses a custom-built electrode pair to compute bio-impedance and Cole parameters covering a wide range of frequencies from 1 Hz to 1 MHz [4]. However, impedance of human cell and tissue consists of resistive and capacitive components [5], [6]. It's determined by injecting an alternating current in the cell or tissue and then measuring the output voltage across it. The Linear Time-Varying (LTV) bio-impedance is measured with a specified precision using stepped-sine excitations, as given in [7], but it is susceptible to temporal distortions affecting the data, which limits the device's temporal bandwidth and sets the data accuracy. Current source and voltage sensing circuit are the essential blocks in the instrument. Several authors have successfully designed current sources operating up to few hundreds of kHz [8]. Paul Annus et.al. [9] have systematically analyzed the design of a current source using transfer function approach and they have measured the load impedance using load in loop method or configuration. The voltage sensing circuit consists of amplifier, demodulator and low pass filter. The use of instrumentation amplifier for bio-impedance measurements has been analyzed by Areny and Webster [10]. The measurement technique has been used in a

number of applications such as calculating Total Body Water (TBW), calculating Intracellular Fluid (ICF) and extracellular fluid (ECF) [6] Electrical Cardiometry [11], Skin Water Content, Impedance Imaging (Tomography), Ablation Monitoring and measurement of Respiration Rate [12]. The technique also has the potential to be used in biometrics [13]. Body Composition Assessment, Transthoracic Impedance Pneumography, Electrical Impedance Tomography (EIT), and Skin Conductance are examples of bio-impedance applications that are described and analysed in [14]. [15] looked into the possibility of non-invasively tracking blood glucose levels using bio-impedance data, which would allow for more regular testing and better diabetes management and monitoring. The bio-impedance measurement is not a new technique in biomedical diagnostic techniques but research is still going on to make this technique a standard procedure for diagnosis of a particular disease. For this purpose, it is required that a large amount of data be collected for a particular disease, for a particular environmental society and analysis to be done, correlating the disease with the signal recorded. To attain this objective, we have designed a low-cost instrument using common analog signal processing blocks, which gives an accurate recording of the bio-impedance signal.

### III. BACKGROUND

#### A. Basics and Origin of Bio-impedance

Bio-impedance is a passive electrical property that describes a biological cell or tissue's ability to obstruct (oppose) the flow of electrical current through it. The reaction to electrical excitation (current or potential) applied to biological tissue is used to determine bio-impedance. The same or other electrodes applies the excitation signal and picks up the reaction in bio-impedance measurements, then charge conversion takes place from electronic to ionic charge and vice versa [7]. Simply, the ratio of voltage (V) to alternating current (I) is known as electrical impedance (Z). Since, Direct Current (DC) is quite hazardous to humans, therefore, it is never used for any experimentation on humans. In fact, Alternating Current (AC) is more preferable choice for such type of applications. The calculated or observed bio-impedance (Z) is highly influenced by the Resistive (R), Capacitive (C), and Inductive (L) parts of the cells and tissues. The bio-impedance (Z) is given by using the modulus  $|Z|$  and the phase change. Since, bio-impedance (Z) is a complex function or parameter, therefore its Resistance (R) is the real part and whereas, the Capacitive Reactance ( $X_c$ ) is the imaginary part. As the ICF, CM, and ECF are made of dissimilar materials with nonidentical electrical properties, therefore every cell and tissue components react differently to the applied AC signal. As we know that ICF and ECF are made up of ionic solution which is highly conducting in nature, thus it provides low resistance path to the applied AC signal [16]. The CM are composed of lipid bilayers which are electrically nonconducting and inserted between two layers of conducting proteins. This sandwiched structure, produces a capacitive reactance ( $X_c$ ) to the applied AC signal [17], [18]. Due to this, biological cells and tissues produces a complex bio-impedance (Z) which can be considered as overall response to an applied AC signal [2], [3]. Thus, bio-impedance (Z) is a complex function depends upon cell and tissue composition and structure, health of person and applied AC signal frequency. Moreover, it also changes with measurement direction,

from one subject to the other subject and even within the tissue itself.

The human body composition comprises, water (64%), protein (20%), fat (10%), and minerals (5%) and starch (1%). The human body mass is mainly due to O (65%), C (18%), and H (8%). The majority of muscles are made up of protein whereas, majority of bones are made up of minerals [19]. The bio-impedance is proportional to Total Body Water (TBW), which contains Intra-Cellular Water (ICW) and Extra-Cellular Water (ECW). Body water, body fat, and body muscle have different impedance values according to the amount of presence of water in these. Thus applied AC signal pass through paths that contain more water as it provides high conductivity [20]. The physiological, morphological, pathological settings and also applied AC signal frequency, all these affect and influence cells and tissues and their electrical properties [21], [22]. The biological cells and tissues may have active (endogenous) or passive (exogenous) electrical properties, depending on the type of source of applied AC signal. The bio-electric signals from the heart known as electrocardiograph (ECG), signals from the brain known as electroencephalograph (EEG), whereas electromyograph (EMG) signals from the muscles are few examples of active properties (bio-electricity) produced by ionic activities within cells and tissues (typical of nerve cells). Passive properties are generated by simulating them with an external electrical excitation source [23], [24].

The extracellular fluids surround all cells with membranes in biological tissues. The main constituents of Extra-Cellular Fluid (ECF) are fluid component of the blood known as plasma and the other one is Interstitial Fluid (IF) which surrounds all cells that are not in blood. The extracellular space is the part of a multicellular organism outside the cells, whereas intracellular space is within the organism's cells. The cell membranes separate extracellular spaces and intracellular space thus producing two electrically conducting compartments known as extracellular media and intracellular media. The resistive pathways are provided by ECF and intracellular fluids (ICF). Due to its insulating design and structure, the lipid bilayer cell membrane is very-very thin measuring approximately 6-7 nm. This lipid bilayer cell membrane is semi-permeable, due to which it has a high capacitance and which produces capacitive reactance [25], [26], [27]. Although biological cells and tissues may have inductive properties, inductance is much lower at low frequencies than resistance and reactance, so it is often overlooked [28]. Thus, biological cells or tissue's complex bio-impedance is the contributions from both frequency-dependent capacitance and conductance [21], [29], [30], [31], [32], [33]. Bio-electrical impedance often differs from one tissue to the next, as well as from one subject to the next. The complex bio-electrical impedance is affected by changes in cell and tissue composition and structure, and even health condition or status of the subject [5], [6].

#### B. Frequency Response of Bio-impedance

The anatomical, physiological, and pathological state of biological cells and tissues determine the bio-impedance frequency response. Therefore, the bio-impedance study can provide much more information related to the anatomy and physiology of a cell or tissue. Since the bio-impedance response is a variable of signal frequency, therefore bio-impedance analysis

with multifrequency inputs can give detailed cell or tissue attributes information, that can help greatly in cell or tissue specifications. Furthermore, the applied ac signal frequency also drastically affects it [2], [3]. Thus, the bio-impedance of cells or tissues and their frequency response is greatly affected not only by the composition and structure of these cells and tissues in the given physiological and physiochemical setting but also by the applied signal frequency. The few bio-impedance analysis techniques which makes use of lumped estimation of the bio-impedance values of the cell or tissue samples are BIA, IPG, and ICG etc. Bio-Electrochemical Impedance Spectroscopy (EIS) measures and analyses bio-impedance at different frequencies. Thus, EIS provides not only a lumped approximation of the cell or tissue sample's bio-impedance values at relatively higher frequency (generally 50 kHz), but also the details required to have better understanding of the many complex bio-electrical phenomena such as dielectric dispersions and relaxation.

The bio-impedance measurement can be broadly divided into two categories, namely, "single-tone" signals and "multi-tone" signals measurements. The analysis of "single-tone" signals is very straightforward, but measurements take longer, whereas use of a multi-tone signal allows for simultaneous coverage of the entire frequency spectrum. However, use of a multi-tone signal may result in an algorithm which can be more complex for analysis purpose [34]. Moreover, especially the dielectric properties of the object determine the required frequency range for bio-impedance measurements, which typically spans 3 to 4 decades in the kHz to MHz range. In general, wider range of frequencies improves fitting accuracy but at the cost of complicating measurements. Furthermore, (SNR) of measured signals also effects the fitting accuracy [35].

### C. Types of Electrode Configurations

When an alternating current is used in bio-impedance testing, the electrode displays a frequency-dependent Electrode Polarisation Impedance (EPI) at the contact point with the tissue or solution as the case may be. Any change in the electrode material that is in contact with the tissue or solution also affects the magnitude and as well as the phase of the electrode impedance. As a result, the total calculated impedance of the system is equal to the sum of the EPI and impedance of the tissue/solution [25]. Finally, the impedance depends upon type of electrode used, electrode material used, the applied signal amplitude and also on its size, shape and structure [36]. In order to measure bio-impedance, we need at least two electrodes for the electrical current to pass through the closed circuit [25]. Therefore, bio-impedance measurements are performed with two or four electrodes. In both methods, the red coloured electrodes depicted in Fig. 1 are known as input or excitation electrodes. Basically, these are the current or driving electrodes. On the other hand, blue coloured electrodes also depicted in Fig.1 are known as voltage/sensing or output electrodes used to determine output signal which is frequency dependent. It should be observed that bio-impedance measurements can be inaccurate due to factors such as movement and improper electrode placement [36]. Bio-impedance measurements are typically done with gel electrodes to minimise electrode-skin impedance. The usability of dry electrodes is studied in [37] because this type of electrode

is not suitable in many measurement environments. For bio-impedance measurement, there are two kinds of electrode configurations. According to its name the two-electrode system or setup depicted in Fig.1(a) for measurement of impedance makes use of two electrodes only. Thus, the current signal injection or current-carrying as well as voltage measurement or voltage pickup are done with the same electrodes. For unipolar measurements, electrode configuration with quite large and small electrodes can also be used [38].

The two-electrode technique suffers from contact impedance due to polarisation impedance at the electrodes' surfaces and the measured signal also covers the contact impedance due to voltage drop [25]. While, analysing the measured signal, the polarisation impedance should be considered and removed in the output signal [39], [40]. Overall, the results obtained using this method are interesting, however they do not provide an accurate signal on the electrode's surface [41]. In four electrode configuration or setup, two independent pair of electrodes are used for current injection and detecting changes in voltage or voltage measurements [39], [42]. In this system, the input that is constant amplitude current signal is injected using the outer electrodes. These electrodes are also known as current or driving electrodes depicted in red colour in Fig.1(b). The output voltage signal produced which is frequency dependent is measured at two points within the current electrode. These electrodes are also known as voltage or sensing electrodes depicted in blue colour in Fig. 1(b). In this configuration, as the distance between electrode pairs was increased, the magnitude of the measured impedance decreased [25], [43]. A number of factors influence the effect of electrode polarisation impedance which includes electrode content, size, measuring frequency, sample impedance, and so on. The main advantage of four electrode configuration over its counterpart two electrode configuration is that the voltage or sensing electrodes do not carry current, due to this it eliminates polarisation impedance. Thus, at the connecting surface of electrode with tissue or electrolyte the influence of contact impedance is greatly decreased. Therefore, use of this method is a common and popular practice to lower the effect of EPI [44]. Furthermore, four electrode configuration measurements are more sensitive and accurate.

## IV. PROPOSED METHOD

The instrument designed consists of power supply, current source, voltage sensing unit and data acquisition system (DAS). Fig.2 depicts the general block diagram of the measuring instrument. The current source circuit is used to generate a sinusoidal current of amplitude in the range of 600  $\mu$ A- 800  $\mu$ A and a frequency of 50 kHz [4]. The voltage sensing unit is used to amplify and remove the high frequency components from the signal sensed by voltage sensing electrodes. The signal is then given to the data acquisition system. The contact surface dimensions of all test electrodes are the same, and the carrier is a circular printed circuit board. To compare the electrodes' characteristics, the electrode-skin impedances are measured under a variety of signal frequencies, contact durations, contact pressures, positioning positions, and subjects. All measurements are often done with silver/silver chloride (Ag/AgCl) dry gel electrodes for contrast. [37] The instrument has four Ag/AgCl electrodes, two on the outside and two on the inside,

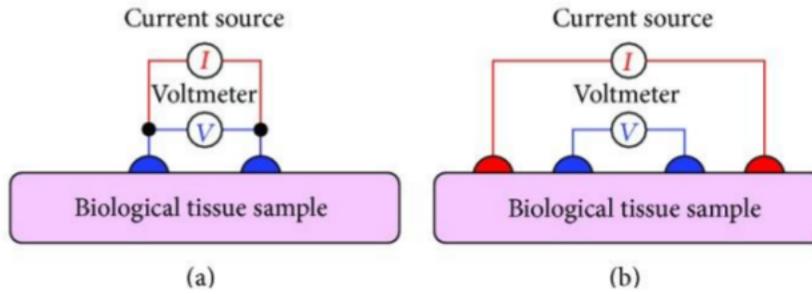


Fig. 1. Bio-impedance Measurement using: (a) Two- Electrode Method, (b) Four-electrode Method.

which are used as current electrodes and voltage electrodes, respectively. This type of arrangement is known as a tetra polar arrangement. Such an arrangement has been implemented by J. J. Wang et.al. [45] to use forearm impedance plethysmography for monitoring cardiac pumping function.

#### A. Waveform Generation

This is the first stage of the instrument and also the most critical one. All the parameters for this stage are very important as far as subject's safety and the output of the device is concerned. As this is the first stage any noise added at this stage will get amplified in the subsequent stages. A good waveform generator should have a single frequency output with very low value of total harmonic distortion and should have a very low frequency drift. A Wein bridge oscillator can be used to generate a sinusoidal waveform [46]. The oscillator is designed for a frequency of 50 kHz as this frequency is widely accepted in clinical use as a standard [47]. Among all types of voltage source generators, an oscillator produces the output with the most stable frequency as compared to other sources. Wein bridge oscillator is implemented using an op-amp, which gives a stable output with a single frequency for a particular combination of resistors and capacitors. Proper limiter circuit is also used in the circuit to keep the poles of the oscillator on the imaginary axis. One major problem with using Wein bridge oscillator is that the instrument designed cannot be used for bio-impedance spectroscopy.

A comparator, integrator and wave shaping circuit can be used to generate square, triangular and sinusoidal waveform of variable frequency [48]. The comparator and integrator circuit used together generate square and triangular waveforms. Triangular waveform is shaped using a shaping circuit to generate sinusoidal waveform. The maximum frequency achieved by this circuit using 741 op-amp is 26 kHz. This method can be used for excitation in lower frequency range but lower frequencies are not used in bio-impedance spectroscopy as electrodes get polarized at lower frequencies. It has also been observed during our laboratory experiments that there was no biomodulation obtained in the output signal when a current of frequency less than 20 kHz was injected into human body. The instrument designed uses a monolithic integrated circuit ICL8038, to produce high accuracy sine, square and triangular waveforms [49]. The frequency can be selected externally from 10 kHz to 100 kHz using a potentiometer. The output of ICL8038 is stable over a wide range of temperature and supply variations. The total harmonic distortion for ICL8038 varies

from 1% to 2% depending on the model selected.

#### B. V To I Convertor

The current driver is one of the most important sub circuit for the measurements of bio-impedance. The current driver can easily work over a fairly wide range of impedance and frequency. The main requirements of a current driver are high output impedance, short phase delay and minimal harmonic distortion. Depending on whether they are open loop or closed loop, these are categorized into two groups. The features of each design are described [50].

The voltage waveform generated using 8038 is converted into current, which is injected into human body. The current generated is within the safe limits ( $600\mu\text{A}$ -  $800\mu\text{A}$ ). The V to I converter is intended to run in the frequency range of 10-100 kHz.

The main requirement for a current source is that it should supply a constant amount of current irrespective of the impedance of the load connected to it. When the output impedance of the current source is much greater than the load impedance then the current through the load is maintained constant regardless of the load value. F. Seoane et.al have analyzed Howland circuit as V to I converter [51]. In this instrument Howland circuit with buffer is used for voltage to current conversion. Fig. 3 is the circuit diagram for voltage to current converter. The buffer stage increases the output impedance of the V to I converter. The input impedance of the circuit is around 20 k $\Omega$ . Fig. 4 depicts the simulation results for circuit of Fig.-3. The resistance {R} in the Fig. 3 represents the load to which current would be injected. The human body impedance is in the range 1 k $\Omega$  - 3 k $\Omega$ . In simulation the resistance {R} is varied from 0.1k to 5k in steps of 100  $\Omega$ . It is verified using the simulation results that the current through the load is almost constant (10  $\mu\text{A}$  variation) irrespective of the load resistance.

#### C. Instrumentation Amplifier

The voltage sensed from the section of human body is in range of millivolts. The input signal is amplified using a difference amplifier. A single op-amp can also be used as a difference amplifier but due to its low input impedance it cannot be used. The input impedance of the single op-amp difference amplifier can be increased by introducing a buffer at each of the inputs of the amplifier and instead of using a unity gain follower one can also have some gain from the first

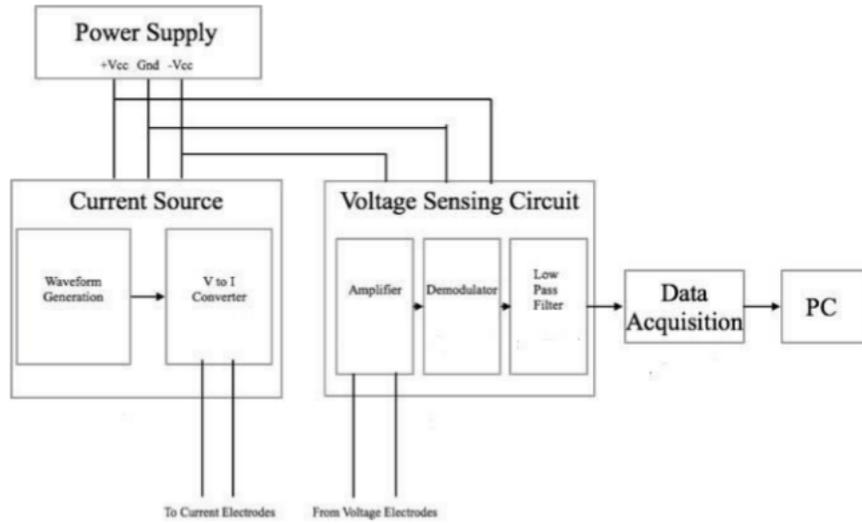


Fig. 2. General Block Diagram of Measuring Instrument.

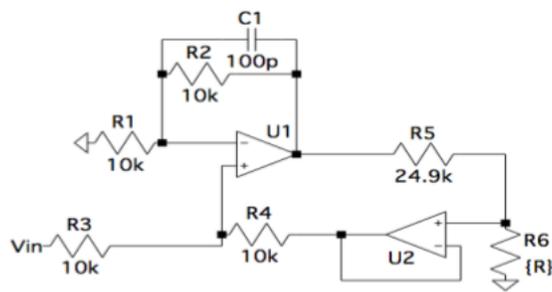


Fig. 3. Circuit Diagram for Voltage to Current Converter.

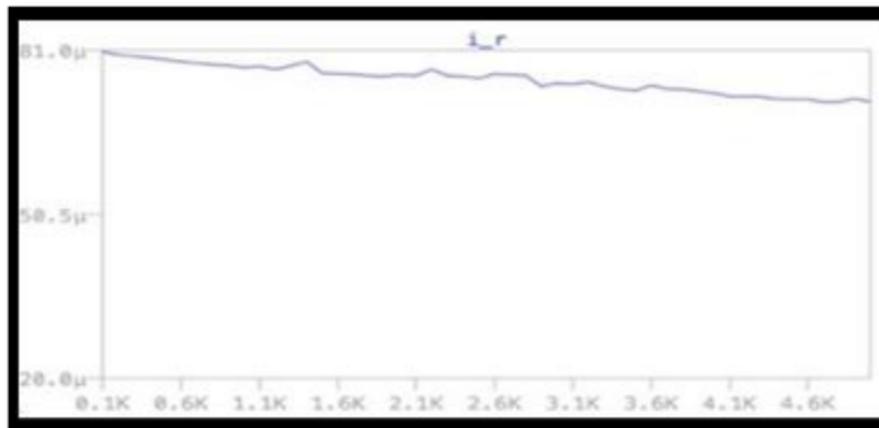


Fig. 4. Variation of Current Through {R} with Respect to Change in Value of Load Resistance, y-axis is the Current Through the Load and x-axis is the Impedance of the Load Connected to the V to I Converter.

stage. This leads to the circuit of instrumentation amplifier as shown in Fig. 5. The instrumentation amplifier has high CMRR and high input impedance and has been very effectively used in biomedical applications such as EEG and ECG.

The analysis for instrumentation amplifier has been given by Adel S. Sedra and Kenneth C. Smith [1] and its use as a biopotential amplifier has been analyzed by Nagel J.H [52]. The gain for the instrumentation amplifier is given below.

$$\text{gain} = \frac{R_4}{R_3} \left( 1 + \frac{R_2}{R_1} \right) \quad (1)$$

The signal from the electrodes is given as input to the instrumentation amplifier through its two terminals. The gain of the amplifier is 26.44 dB. The first two op-amps are used for gain and the third operational amplifier is used for common mode rejection. In the circuit designed  $R_4=R_3$  and  $R_2/R_1$  is equal to 20. The amplifier's input impedance is equal to the input impedance of the operational amplifier. Frequency response analysis of the instrumentation amplifier depicts that the phase and gain are constant in the desired frequency range. Fig. 6 depicts the frequency response of the instrumentation amplifier in which the continuous plot is the magnitude plot and dotted plot is the phase plot.

#### D. Demodulator

The signal obtained from the human body is an amplitude modulated signal, where the carrier is the current waveform that we have introduced into the human body. In this case the high frequency carrier is a sinusoidal waveform with a frequency of 50 kHz and the modulating bio-impedance signal is low frequency signal with frequency components less than 50 Hz. Webster and Tompkins [53] have suggested use of full wave rectifier for demodulation. The modulating signal can be extracted using a simple envelope detector circuit as shown in Fig. 7. Asynchronous demodulation has been used so that square and triangular waveform excitation can also be used in addition to the sinusoidal waveform.

For the first cycle diode is forward biased and it charges the capacitor to the first peak value. The charging time constant should be such that the capacitor voltage follows the input signal.

$$\tau \text{ charging} \ll \frac{1}{f_c} \quad (2)$$

where  $f_c$  is the frequency of the carrier. The charging times constant depends on source resistance  $R_s$ , forward bias diode resistance  $r_d$  and capacitance  $C_1$ .

$$\tau \text{ charging} = ((R_s + r_d + R_2) \parallel R_1) C_1 \quad (3)$$

$$R_s + r_d + R_2 \ll R_1 \quad (4)$$

$$\therefore \tau \text{ charging} = (R_s + r_d + R_2) C_1$$

When the input signal drops, diode becomes reverse bias (as capacitor is charged to a higher voltage) and the capacitor voltage remains at the initial level. During this time (when diode is reverse bias) the capacitor discharges through the resistor  $R_1$ .

$$\tau \text{ discharging} = R_1 C_1 \quad (5)$$

The discharging time constant should be large so that the capacitor discharges slowly but it should not be so large that it is unable to trace the low frequency modulating signal. The discharging time constant should follow the following relation:

$$\frac{1}{B} \gg \tau \text{ discharging} \gg \frac{1}{f_c} \quad (6)$$

Where B is the bandwidth of the low frequency bio-impedance signals. For the calculation of the values of resistances and capacitance, value of B is taken equal to 50 Hz. The diode used in the circuit is forward biased when the incoming voltage is greater than 0.7V, therefore the low voltage signals cannot be detected. Taking into account this problem related to cut off voltage of diode, a precision diode [46] is used in this circuit in place of normal diode. A precision diode is an operational amplifier with a diode in the negative feedback followed by a diode whose anode is connected at the output pin of the op-amp. The cutoff voltage for a precision diode is approximately equal to zero volts. Fig. 8 is the circuit for precision diode. The output of the demodulator is given to the low pass filter stage through a buffer. Improved output buffering and peak detector gain greater than unity is achieved with an output voltage follower. This leads to circuit of precision envelope detector in Fig. 9. The precision envelope detector is much more accurate than the simple envelope detector as the voltages below 0.7V are also detected by this circuit. Droop due to detector diode leakage can be removed through the use of bootstrapping feedback that holds detector diode bias at zero when the diode is not conducting.

#### E. Low Pass Filter

The signal from the demodulator contains high frequency components, which needs to be filtered out before the signal is given as input to the analog to digital converter of data acquisition. Here, an antialiasing filter is used as a Low Pass Filter (LPF). Accordingly, a second order Chebyshev LPF with a cutoff frequency of 40 Hz is used. Chebyshev filters differ from Butterworth filters in that they have a roll-off which is steeper and more ripple in passband (type I) or stopband (type II). Chebyshev filters have the property of minimizing the difference between real filter characteristics and the idealized one over the desired filter range, but with passband ripples. The LPF output is given to data acquisition system implemented using Arduino uno board.

#### F. Assumptions, Measurement Protocol and Data Acquisition

According to the BIA assumptions, the human body can be considered as a homogenous conductor having cylindrical dimensions (Fig. 10). In which bio-impedance is directly proportional to 'L' and inversely proportional to 'A', where 'L' is the cylinder's length and 'A' is the cylinder's base cross-sectional area (Fig. 10).

BIA formulation processes typically make the following assumptions for ease of calculation, though in practice the human body varies from these assumptions:

- The human body is considered as cylindrical.
- The cylindrical shape is defined by its height and weight.

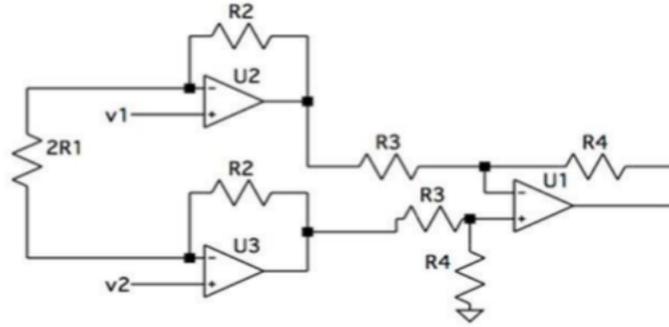


Fig. 5. Circuit Diagram for Instrumentation Amplifier.

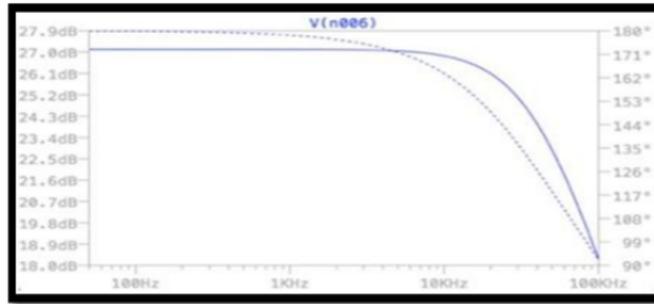


Fig. 6. Frequency Response Analysis of the Instrumentation Amplifier.

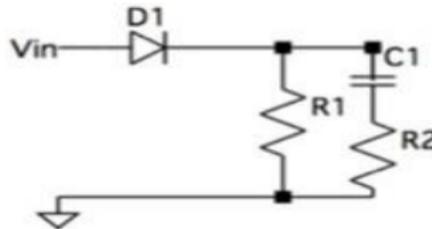


Fig. 7. Circuit Diagram of Simple Envelope Detector.

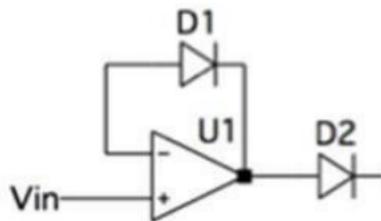


Fig. 8. Precision Detector.

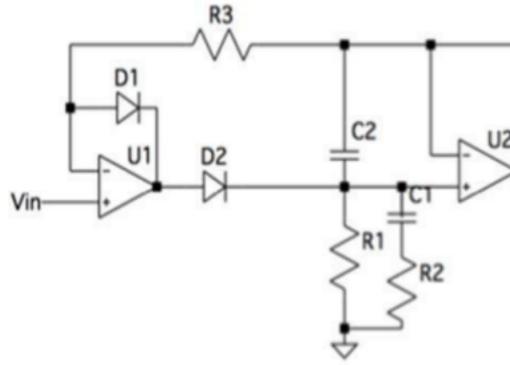


Fig. 9. Precision Envelope Detector.

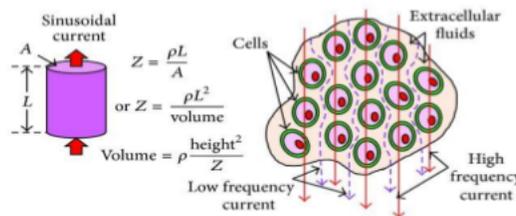


Fig. 10. The Human Body's Impedance when it is Modelled as a Homogeneous Conductor with Cylindrical Volume.

- Homogeneous and evenly distributed body composition is considered.
- The body compositions have no individual differences or variations.
- The environment parameters like temperature, other physiological parameters such as body heat or stress are assumed to be constant.

The cell membranes produce capacitive reactance due to the capacitive nature and caused by a selectively applied AC signal allows current to pass through these cell membranes using current paths largely depends upon the signal frequency (Fig. 10). The cell membranes are penetrated by current with high frequency. Thus, the current penetrates ECFs, CMs and ICFs. The low frequency current passes through ECF only as the cell membrane reactance prevents its flow through it. The BIA technique can be used to determine the (TBW) which is nothing but combination of ECW and ICW. However, it should be done at a particular frequency of the applied AC signal.

The instrument designed was tested on patients at St. Joseph hospital, Ghaziabad, India. Verbal prior approval for the study was taken telephonically from the concerned authorities. Twenty-five adult human subjects (aged 25-60 years) participated in this study. The general test setup for BIA is shown in Fig. 11. However, in our study it differed little bit for the ease of the subjects. They were made to rest in supine position for about ten to fifteen minutes. Then, Ag/AgCl electrodes were put their right and left forearms. Volunteers were told to lie down straight and remain still during measurement.

## V. RESULTS

The test setup used to validate the instrument is similar to the setup used in impedance plethysmography for the cardiac output measurement. In general, impedance is a measure of resistance. Plethysmography has become the gold standard for measuring changes in blood volume in any part of the body based on electrical impedance changes [54]. It is also being used in the diagnosis of peripheral vascular diseases. In impedance plethysmography, values of instantaneous impedance ( $Z$ ), basal impedance ( $Z_0$ ), which is an average of calculated bio-impedance values and derivative of impedance with respect to time ( $dZ/dt$ ) is used to calculate cardiac output parameters. Since the technique is a standard procedure, the values of basal impedance measured by this technique can be used to validate the instrument designed. The output of the demodulator is essentially the basal impedance of the section across which electrodes have been applied. The basal impedance values of forearm given in literature, [55] are used to validate the output signal of the instrument designed.

The output voltage signal of three sets of volunteers has been displayed in Fig. 12. The slight variations in the voltage values depict the variations in impedance ( $Z_0$ ). The basal impedance is calculated using the constant voltage level. The calculated value of basal impedance and its comparison with values of basal impedance in literature has been shown in Table I. The comparison in Table I shows that the values measured from the output of the instrument is in accordance with that of standard technique, which validates the design of the instrument.

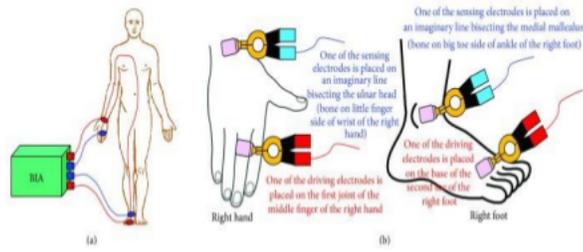
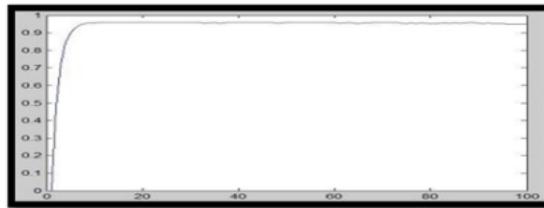


Fig. 11. General Test Setup for BIA.

TABLE I. BASAL IMPEDANCE (FOREARM)

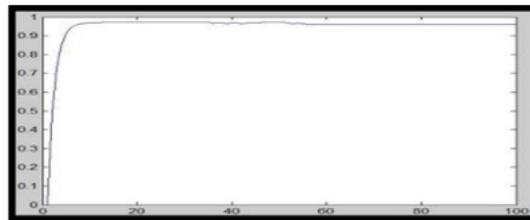
| SET OF VOLUNTEERS [NOs] | MEASURED BY INSTRUMENT DESIGNED |                         | STANDARD VALUES IN LITERATURE |                         |
|-------------------------|---------------------------------|-------------------------|-------------------------------|-------------------------|
|                         | $A_V \cdot AGE$                 | $A_V \cdot Z_0(\Omega)$ | AGE GROUP                     | $A_V \cdot Z_0(\Omega)$ |
| A [5]                   | 40                              | 58.16                   | 36-45                         | $65.44 \pm 12.03$       |
| B [8]                   | 48                              | 80.57                   | 46-55                         | $67.50 \pm 6.38$        |
| C [12]                  | 57                              | 58.77                   | > 55                          | $69.13 \pm 8.74$        |



(a)



(b)



(c)

Fig. 12. Output Voltage Signals Measured using the Instrument. (a) is the Output Signal of Volunteer Set-A (b) being the Output Signal of Volunteer Set-B and (c) is the Output Signal of Volunteer Set-C. The y-axis in Each Figure is the Output Voltage Amplitude Value and the x-axis Represents Number of Samples.

## VI. CONCLUSION

A simple low cost bio-impedance measuring instrument has been designed using common electronic blocks. The results have been validated and have been found to be accurate and reliable. Each block of the instrument has been tested using simulation and verified experimentally. The method and device developed fulfills required specifications and can be used in clinical examinations. Bio-impedance measurement techniques can prove very useful for first hand diagnosis. The major problem for using this technique for diagnosis is that its results are not standardized. By standardized one means that just by looking the graph or using its results one can say whether the results correspond to a normal person or a patient suffering from a disease. A large amount of analysis has to be done for a particular application on bio-impedance signals to enable the doctors to rely on results produced by bio-impedance analysis. The standardization of bio-impedance values for human body may be difficult due to variable ambient conditions, eating habits, lifestyle and anthropological background. So intensive and extensive research and in-depth analysis is required to be carried out towards normalization and standardization of the bio-impedance values.

Researchers in the field of bio-impedance have discovered that multifrequency bio-impedance analysis (BIA) is one of their main interests [56]. A group of researchers are exploring instrumentation and designing more advanced instrumentation. In the future, investigation and research can be carried out for Bluetooth-enabled wireless instrumentation for BIA techniques. Moreover, in future studies Bluetooth-enabled wireless instrumentation for Electrochemical Impedance Spectroscopy (EIS) techniques could be investigated. To improve cardiac health monitoring and to have potential transthoracic parameter assessment; ICG along with a bio-impedance tool with multifrequency features can be used. Furthermore, in an ambulatory or long-term monitoring requirement in the Intensive Care Unit (ICU) bio-impedance based ICG can be used. However, if future research on Bio-impedance calculation and BIA can overcome these challenges, it can be implemented more appropriately than other commonly used and popular imaging methods in specific medical diagnostic applications especially related to the imaging of brain, breast, abdominal and whole body, etc.

Bio-impedance technology is becoming more prevalent as a result of an increasing number of healthcare monitoring applications. The study presented has a wide range of implications, including the ability to improve bio-impedance studies and healthcare devices that use bio-impedance technology [7]. Bio-impedance spectroscopy measures the bio-impedance of cells and tissues covering fairly good frequency range. Physiological testing and health-monitoring systems benefit greatly from this method. Devices must be compact, wearable, or even implantable for a wide variety of applications. As a result, the next generation of bio-impedance sensing systems must be designed to save energy and resources [34]. It's critical to consider the application as well as the type of cell or tissue culture to be monitored when selecting a bio-impedance measurement technique [57]. As a result, choosing the right electrode configuration is crucial. Future research should focus on the electrodes and the bio-impedance measurement method [57]. Overall, this paper describes a low-cost, bio-electrical impedance

measurement system that has been successfully developed and tested and can be used effectively and efficiently for non-invasive health monitoring. The paper also discussed some of the most important technological aspects and limitations of bio-impedance calculation and study. Finally, in this paper the theoretical dimensions, operating principles, implementations, benefits, disadvantages, and current research status, upcoming developments, and bottlenecks in bio-impedance analysis and calculation all have been covered in great detail.

## STATEMENTS & DECLARATIONS

### FUNDING

The authors hereby declare that no funds, grants, or other support in any form were received during the preparation of this manuscript.

### COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

### AUTHOR'S CONTRIBUTION

All authors contributed almost equally for material preparation, data collection and analysis to the said study related to conception and design of bio-impedance measuring instrument. Moreover, all authors read and approved the final manuscript.

### ETHICS APPROVAL

Being an observational study only, it did not require any formal ethical approval. However, verbal consent from the competent authority obtained telephonically.

### CONSENT TO PARTICIPATE

No such Informed consent was required for this observational study, however individual participants if any were informed verbally.

### CONSENT TO PUBLISH

As no figure in the manuscript depicts human participants, therefore no such consent for publication of the images of the human research participants is required.

## REFERENCES

- [1] F. H. Netter, "Atlas of Human Anatomy, Rittenhouse. 2nd. Book Distributors Inc," 1997, [Google Scholar].
- [2] J. J. Ackmann, "Complex bioelectric impedance measurement system for the frequency range from 5 Hz to 1 MHz," *Annals of Biomedical Engineering*, vol. 21, no. 2, pp. 135–146, 1993, [PubMed] [Google Scholar].
- [3] K. Cha, G. M. Chertow, J. Gonzalez, J. M. Lazarus, and D. W. Wilmore, "Multifrequency bioelectrical impedance estimates the distribution of body water," *Journal of Applied Physiology*, vol. 79, no. 4, pp. 1316–1319, 1995, [PubMed] [Google Scholar].
- [4] A. Roy *et al.*, "An experimental method of bioimpedance measurement and analysis for discriminating tissues of fruit or vegetable," *AIMS Biophysics*, vol. 7, no. 1, pp. 41–53, 2020.
- [5] S. Grimnes and G. Martinsen, "Academic Press, Bioimpedance & Bioelectricity Basics," 2021.

- [6] L. Jodal, "MSc, Medical Physicist, lecture notes on the electrical theory behind the measurement of body fluids with bioimpedance spectroscopy (BIS)," 2021.
- [7] B. Louarroudi and Sanchez, "On the correct use of stepped-sine excitations for the measurement of time-varying bioimpedance," *Physiological Measurement*, vol. 38, no. 2, pp. N73–N80, 2017.
- [8] J. W. Haslett and M. K. N. Rao, "A High Quality Controlled Current Source," in *Instrumentation and Measurement*, *IEEE Transactions on*, vol. 28, no. 2, pp. 132–140, 1979.
- [9] P. Annus, A. Krivoshei, M. Min, and T. Parve, "Excitation Current Source for Bioimpedance Measurement Applications: Analysis and Design," *Instrumentation and Measurement Technology Conference Proceedings*, pp. 848–863, 2008.
- [10] R. Pallas-Areny and J. G. Webster, "AC Instrumentation," 2021.
- [11] "Electrical Cardiometry" Wikipedia, 2021.
- [12] "Respiration rate measurement based on impedance pneumography Application report, SBAA1811-February 2011, Texas Instruments," 2021.
- [13] O. G. Martinsen, "Utilizing characteristic electrical properties of the epidermal skin layers to detect fake fingers in biometric fingerprint systems-A pilot study," *Biomedical Engineering*, vol. 54, pp. 891–894, 2007.
- [14] D. Naranjo-Hernández *et al.*, "Fundamentals, Recent Advances, and Future Challenges in Bioimpedance Devices for Healthcare Applications," *Journal of Sensors*, Hindawi, 2019, Article ID 9210258.
- [15] P. S. H. Jose *et al.*, "A Non-Invasive Method for Measurement of Blood Glucose using Bio Impedance Technique," in *2nd International Conference on Signal Processing and Communication (ICSPC)*, 2019, IEEE.
- [16] J. Nyboer, "Electrical impedance plethysmography; a physical and physiological approach to peripheral vascular study," *Circulation*, vol. 2, no. 6, pp. 811–821, 1950, [PubMed] [Google Scholar].
- [17] R. W. Griffiths, M. E. Philpot, B. J. Chapman, and K. A. Munday, "Impedance cardiography: non-invasive cardiac output measurement after burn injury," *International Journal of Tissue Reactions*, vol. 3, no. 1, pp. 47–55, 1981, [PubMed] [Google Scholar].
- [18] C. J. Schuster and H. P. Schuster, "Application of impedance cardiography in critical care medicine," *Resuscitation*, vol. 11, no. 3-4, pp. 255–274, 1984, [PubMed] [Google Scholar].
- [19] "Internet Article, JAWON Medical, South Korea," 2021, <http://www.jawon.co.kr/eng/technology/body-composition/principles-of-bia.php>.
- [20] "Internet Article. Bodystat Limited, USA," 2021, <http://www.bodystat.com/science>.
- [21] D. Miklavcic, N. Pavselj, and F. X. Hart, "Electric Properties of Tissues," *Wiley Encyclopedia of Biomedical Engineering*, 2006, [Google Scholar].
- [22] H. P. Schwan, "Electrical properties of tissue and cell suspensions: Mechanisms and models," *Proc IEEE Adv Biol Med Soc*, vol. 1, pp. 1:A70–A1, 2002, [Google Scholar].
- [23] R. Pethig and D. B. Kell, "The passive electrical properties of biological systems: Their significance in physiology, biophysics and biotechnology," *Phys Med Biol*, vol. 32, no. 933, 1987, [PubMed] [CrossRef] [Google Scholar].
- [24] U. G. Kyle, I. Bosaeus, D. L. AD, and A. D, "Bioelectrical impedance analysis part I: review of principles and methods," *Clinical Nutrition*, vol. 23, pp. 1226–43, 2004, [PubMed] [CrossRef] [Google Scholar].
- [25] S. Grimnes and Ø. G. Martinsen, "Bioimpedance & Bioelectricity Basics. Elsevier Science," 2014, 3rd ed. [Google Scholar].
- [26] K. Asami, "Characterization of heterogeneous systems by dielectric spectroscopy," *Prog Polym Sci*, vol. 27, pp. 1617–59, 2002, [CrossRef] [Google Scholar].
- [27] K. Heileman, J. Daoud, and M. Tabrizian, "Dielectric spectroscopy as a viable biosensing tool for cell and tissue characterization and analysis," *Biosens Bioelectron*, vol. 49, pp. 348–59, 2013, [PubMed] [CrossRef] [Google Scholar].
- [28] P. J. Riu and C. On, "Bioelectrical parameters of the whole human body obtained through bioelectrical impedance analysis," *Bioelectromagnetics*, vol. 25, pp. 69–71, 2004, [PubMed] [CrossRef] [Google Scholar].
- [29] C. Gabriel, S. Gabriel, and E. Corthout, "The dielectric properties of biological tissues: I. Literature survey," *Phys Med Biol*, vol. 41, pp. 2231–49, 1996, [PubMed] [CrossRef] [Google Scholar].
- [30] Ø. G. Martinsen, S. Grimnes, and H. P. Schwan, "Interface phenomena and dielectric properties of biological tissue," *Encyclopedia of Surface and Colloid Science*, vol. 20, pp. 2643–53, 2002, [Google Scholar].
- [31] D. A. Dean, T. Ramanathan, D. Machado, and R. Sundararajan, "Electrical Impedance Spectroscopy Study of Biological Tissues," *J Electrostat*, vol. 66, no. 3-4, pp. 165–77, 2008, [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [32] H. Kwon, A. L. Mcewan, T. I. Oh, A. Farooq, E. J. Woo, and J. K. Seo, "A local region of interest imaging method for electrical impedance tomography with internal electrodes," *Comput Math Methods Med*, vol. 9, 2013, [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [33] J. K. Seo, T. K. Bera, H. Kwon, and R. J. Sadleir, "Effective Admittivity of Biological Tissues as a Coefficient of Elliptic PDE," *Computational and Mathematical Methods in Medicine*, vol. 2, 2013, [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [34] F. Soulier *et al.*, "Very Low Resource Digital Implementation of Bioimpedance Analysis," *Sensors (Basel)*, *MDPI*, vol. 19, p. 3381, 2019.
- [35] J. Ojarand *et al.*, "How many frequencies to use in electrical bioimpedance measurements," in *Impedance Spectroscopy. Advanced Applications: Battery Research, Bioimpedance, System Design*. Publisher: Walter de Gruyter, 2018, pp. 161–168.
- [36] S. F. Khalil, M. S. Mohktar, and F. Ibrahim, "The Theory and Fundamentals of Bioimpedance Analysis in Clinical Status Monitoring and Diagnosis of Diseases," *A Review. Sensors*, vol. 14, 2014, [PMC free article] [PubMed] [Google Scholar].
- [37] R. Kusche *et al.*, "Dry electrodes for bioimpedance measurements - Design, characterization and comparison," *Biomedical Physics & Engineering Express*, *IOP*, vol. 5, no. 1, 2018.
- [38] H. Kalvøy, L. Frich, S. Grimnes, and Ø. G. Martinsen, "Impedance-based tissue discrimination for needle guidance," *Physiological Measurements*, vol. 30, 2009, [PubMed] [Google Scholar].
- [39] A. Yúfera and A. Rueda, "A Method for Bioimpedance Measure with Four- and Two-Electrode Sensor Systems," in *30th Annual International IEEE EMBS Conference Vancouver*. British Columbia, Canada, 2008, [PubMed] [CrossRef] [Google Scholar].
- [40] X. Huang, "Simulation of Microelectrode Impedance Changes Due to Cell Growth," *IEEE Sensors Journal*, vol. 4, pp. 576–83, 2004, [CrossRef] [Google Scholar].
- [41] I. Giaever, "Use of Electric Fields to Monitor the Dynamical Aspect of Cell Behavior in Tissue Cultures," *IEEE Transaction on Biomedical Engineering*, vol. 33, pp. 242–7, 1986, BME. [PubMed] [CrossRef] [Google Scholar].
- [42] T. S. Carvalho, A. L. Fonseca, A. Coutinho, B. Jotta, A. V. Pino, and M. N. Souza, "Comparison of bipolar and tetrapolar techniques in bioimpedance measurement. XXIV Congresso Brasileiro de Engenharia Biomédica – CBEB," 2014, [Google Scholar].
- [43] L. A. Geddes and R. Roeder, "Criteria for the selection of materials for implanted electrodes," *Annals of Biomedical Engineering*, vol. 31, no. 7, pp. 879–90, 2003, [PubMed] [CrossRef] [Google Scholar].
- [44] H. Kalvøy, G. K. Johnsen, Ø. G. Martinsen, and S. Grimnes, "New Method for Separation of Electrode Polarization Impedance from Measured Tissue Impedance," *The Open Biomedical Engineering Journal*, vol. 5, pp. 8–13, 2011, [PMC free article] [PubMed] [CrossRef] [Google Scholar].
- [45] J. Wang, W. Hu, T. Kao, C. Liu, and S. Lin, "Development of forearm impedance plethysmography for the minimally invasive monitoring of cardiac pumping function," *Journal of Biomedical Science and Engineering*, vol. 4, pp. 122–129, 2011, Inter instrument comparison of bioimpedance spectroscopic analyzers.
- [46] A. S. Sedra and K. C. Smith, "Microelectronic Circuits," 2021, 5th edition Oxford International Student addition.
- [47] "Bioelectrical impedance analysis in body composition measurement: National institutes of health technology assessment conference statement," 1994.
- [48] G. B. Clayton, "88 Practical Opamp Circuits," 2021, 5th edition Oxford International Student addition.

- [49] "ICL 8038 datasheet," 2021.
- [50] N. Neshatvar *et al.*, "Analog Integrated Current Drivers for Bioimpedance Applications: A Review," *Sensors (Basel)*, vol. 19, no. 4, p. 756, 2019, Pub Med.
- [51] S. I. Kim and T. S. Suh, "Current Source for Wideband Electrical Bioimpedance Spectroscopy Based on a Single Operational Amplifier," in *World Congress of Medical Physics and Biomedical Engineering*. Springer-Verlag, 2006, 20006COEX Seoul, Korea. Berlin Heidelberg: Springer-Verlag., 2007.
- [52] "Biopotential Amplifiers in The Biomedical Engineering HandBook," 2000, by Joachim H. Nagel edited by Joseph D. Bronzino.
- [53] M. . Qu, Y. Zhang, J. G. Webster, and W. J. Tompkins, "Motion Artifact from Spot and Band Electrodes During Impedance Cardiography," *Biomedical Engineering*, no. 11, pp. 1029–1036, 1986.
- [54] "25 years of impedance plethysmography," 2003, BARC Newsletter No.236.
- [55] D. L. Prajapati, V. Chintan, Parmar, A. Pradnya, Gokhale, B. Hemant, C. J. S. Mehta, and Non, "Invasive Assessment of Blood Flow Index In Healthy Volunteers Using Impedance Plethysmography," *International Journal of Medical and Health Sciences*, 2013, ISSN 2277-4505.
- [56] B. T. Kanti, "Bioelectrical Impedance Methods for Non-invasive Health Monitoring: A Review," *J Med Eng*, p. 381251, 2014, PMCID: PMC4782691, PMID: 27006932.
- [57] M. Amini, J. Hisdal, and H. Kalvøy3, "Applications of Bioimpedance Measurement Techniques in Tissue Engineering," *J. Electrical Bioimpedance*, vol. 9, no. 1, pp. 142–158, 2018.

# Detecting Irony in Arabic Microblogs using Deep Convolutional Neural Networks

Linah Alhaidari, Khaled Alyoubi, Fahd Alotaibi  
Department of Computing and Information Technology,  
King Abdulaziz University,  
Jeddah, 21589, Saudi Arabia

**Abstract**—A considerable amount of research has been developed lately to analyze social media with the intention of understanding and exploiting the available information. Recently, irony has took a significant role in human communication as it has been increasingly used in many social media platforms. In Natural Language Processing (NLP), irony recognition is an important yet difficult problem to solve. It is considered to be a complex linguistic phenomenon in which people means the opposite of what they literally say. Due to its significance, it becomes essential to analyze and detect irony in subjective texts to improve the analysis tools to classify people opinion automatically. This paper explores how deep learning methods can be employed to the detection of irony in Arabic language with the help of Word2vec term representations that converts words to vectors. We applied two different deep learning models; Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM). We tested our frameworks with a manually annotated datasets that was collected using Tweet Scraper. The best result was achieved by the CNN model with an F1 score of 0.87.

**Keywords**—Verbal irony; natural language processing; machine learning; automatic irony detection

## I. INTRODUCTION

Sentiment analysis is known as the extraction and interpretation of opinions expressed in a text written in a natural language on a certain subject [1]. Recently, sentiment analysis and opinion mining became extremely popular due to the increase of social network usage, which led to produce a huge number of texts. Researchers and many organizations became more interested in analyzing such a text type in order to understand human communication better. Hence, irony is a sophisticated form of sentiment expression where people express their opinions in a certain way [2]. Therefore, it has become an important topic in NLP as it flips the polarity of the posts.

According to Cambridge dictionary<sup>1</sup>, irony is described as a figure of speech that means the opposite of what people really say in a way of being humorous. For example, "I love going to the dentist!". Irony can be either situational or verbal. Situational event occurs when naturally expecting something to happen but the opposite take place [3]. While verbal irony is when individuals express words that represent the opposite of what they actually feel and that is the focus of the existing studies in irony detection. Sarcasm is another term that often occur along with Irony. There exists a lack of conformity in the relationships between irony and sarcasm. Some researchers [4]

[5] define sarcasm as a type of irony in which it is directed at an individual, with the purpose to mock. While others [6], [7] see them as a distinct phenomenon and consider sarcasm differ from irony in which sarcasm include an element of ridicule that irony has not.

In our regular everyday communication, we meet people who like to use irony in the conversation. In most cases, we can detect irony depending on the tone, context, facial expressions, and the person's character. However, when ironic posts in social media are in question, recognizing the irony becomes more challenging and complex due to its ambiguous nature, the missing intonation of the person who writes the message, restricted length of the words, the informal language, the use of hashtags, and the context is not always clear.

Irony has been studied by many research fields such as psychology [8], linguistics education [9], and computational science. It is used widely as an indirect negation in order to achieve different communication goals in many situations such as criticizing, make fun of people, and manipulate answers to upsetting questions.

In recent years, social media have become a part of people's everyday modern life. It enables them to share their opinions on different topics along with other matters. Therefore, the appearance of irony in social networks such as microblogs has greatly increased. For this reason, one of the primary motivations behind this research is detecting real intention behind posts accurately and understanding how people feel regarding specific matters can be useful for many applications. It can help in correctly identifying security issues such as threatening posts by verifying whether the threat words are literal or not. Also, it can be helpful in distinguishing figurative language devices that are used in the different social media platforms, product reviews, feedback, etc., and recognizing the ironic negative reviews/post that being misinterpreted as positive. In general, any tool that aims to extract the meaning of a post effectively can benefit from such a property.

Existing work on detecting irony in Arabic language have mainly focused on classical machine learning [10], [11], [12]. While recent research on detecting irony in English language, such as [13], [14], and [15], applied neural network approaches. Hence, the importance of this research lies on adopting neural network techniques to improve recognizing irony in Arabic texts. To the best of our knowledge, there is only one experiment [12] on detecting irony in Arabic using a neural model, namely BiLSTM. Therefore, we propose a framework that learns irony using a convolutional neural network (CNN).

<sup>1</sup><https://dictionary.cambridge.org/>

In addition, we experimented with Bidirectional Long Short-Term Memory (BiLSTM) as well. The approach we applied outperformed the state of the art. The main contributions of this research can be summarized as follows:

- 1) A manually annotated ironic Arabic dataset.
- 2) We believe that this is the first work on using CNN for irony detection in Arabic texts.

This paper is structured as follows: Section 2 presents a brief literature review on irony detection; Section 3 discusses the approach including data generation process, feature extraction, and the proposed method; Section 4 dedicated to the results; lastly, Section 5 concludes the paper.

## II. RELATED WORK

The interest in adopting neural networks to detect irony on social media has been increased. Classical methods, e.g. SVM, depend on feature engineering and manually convert texts into feature vectors prior to the classification task. In contrast, neural networks approaches can automatically grasp the representations of input texts with different levels of abstraction and then use the gained knowledge to perform the classification task. Several deep learning-based methods have been reported for the field of automatic irony detection. One of the early work was by Poria et al. [16] who proposed an architecture based on a pre-trained convolutional neural network to detect sarcasm on balanced and unbalanced datasets. Sentiment, emotion and personality features were extracted and applied to the system. The balanced datasets achieved the highest f1 score of 0.97. Ghosh et al. [17] applied a semantic neural network model to detect sarcasm over social media content. The architectures are composed of two CNN layers followed by two long short-term memory (LSTM) layers and a deep neural network (DNN) layer. The evaluation of the model achieved an F-score of .92. Ilić et al. [18] proposed a deep neural network model that depends on character-level word representations extracted using the Embeddings from Language Models (ELMo). ELMo is a contextualized representation method that uses vectors extracted from BiLSTM. They tested their system on seven datasets obtained from three different data sources. The results yield 0.87 F-score using Twitter dataset. Authors noted that annotating data manually is necessary in order to improve the performance results. Furthermore, transfer learning approaches, i.e., applying the knowledge of an already trained model to a new task, became very popular in many problems including irony detection in recent research. Potamias et al. [14] proposed a transformer based architecture that builds on the pre-trained RoBERTa model and integrated with a recurrent convolutional neural network (RCNN) that uses non-hand crafted features as they argue that overly trained deep learning approach does not need engineered feature step. They used several benchmark datasets that contain ironic, sarcastic, and figurative expressions. The highest performance of the hybrid neural system achieved f1 score of 0.90 by the sarcastic Riloff's dataset [19]. As for Zhang et al. [20], they focused on finding implicit incongruity without depending on explicit incongruity expressions. They used three transfer learning-based techniques to enhance the attention mechanism of RNNs. The applied attention-based Bi-LSTM achieved higher outcomes on the hashtag-labeled corpus compared to

the human-labeled one. The authors discuss that manually-labeled dataset is considerably more difficult than the hashtag-based dataset and that human annotated dataset results in a more accurate prediction for irony in real applications. Gonzalez et al. [21] used a Transformer Encoder (TE) architecture on two corpora; English and Spanish languages. They applied two TE models with and without the sine-cosine positional information: TE-Pos and TE-NoPos. As a result, TE-NoPos outperformed the TE-Pos system. They also studied the affect of the transformer architecture's multi-head self-attention processes on the irony detection topic. Wu et al. [13] proposed a framework based on four layers of BiLSTM with three dense layers. they combined three tasks which includes finding the missing irony hashtags, classifying ironic or non-ironic and detecting the irony types. The system is concatenated with the sentiment and sentence embedding features that improved the performance by achieving an f1 score of 0.70 and 0.49. Huang et al. [22] considered three deep learning models; Convolutions Neural Network (CNN), Recurrent Neural Network (RNN), and Attentive RNN. The Attentive RNN outperformed other models by achieving F1 score of 0.89. They discussed how attention mechanism improved the irony detection performance. Khalifa and Hussein [12] implemented an ensemble of 8 models based on biLSTM network with TF-IDF, topic modeling and word and character counts features on an Arabic tweets. The ensemble achieved the F1 score of 0.82. Golazizian et al [15] employed a bidirectional LSTM (BiLSTM) network to detect irony in Persian language. They used emoji prediction to construct a pre-trained model that include an attention layer that improved the performance. They reached an accuracy of 83.1. Ren et al. [23] employed two context-augmented neural network methods on Twitter dataset to recognize sarcastic signs from contextual data. Baruah et al. [24] used BiLSTM and BERT transformer based architecture to detect sarcasm in Twitter texts. They applied historical conversational features such as response only and response with varied number of utterances from the dialogue. The F-score of 0.74 was achieved using the BERT classifier. A recent study by Razali et al. [25] which used CNN to extract lexical and contextual features. They chose FastText as word embedding technique. Authors claim that manually extracted contextual features improves the overall accuracy.

## III. APPROACH

### A. Data Generation

A critical challenge for automated irony detection is the availability and quality of a set of ironic examples in order to train a model. This step is performed in three stages shown in Fig. 1.

### B. Data Collection

For the ironic data, we initially collected 12700 Arabic tweets using Tweet Scraper [26]. We gathered the data using the following hashtags: #irony, #Sarcasm, #بتدقيق، #تهكم، #تريقه، #أيروني، #سخريه، #استهزاء

As for the non-ironic datasets, we decided to use a random sample of an existed Arabic sentiment corpus [27].

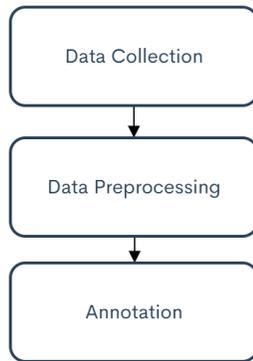


Fig. 1. Data Generation Stages.

### C. Preprocessing of Data

This step is considered as an essential task in sentiment analysis. The goal of preprocessing is to remove corrupt and irrelevant information from raw text. The stream of data we gathered from Twitter is noisy as it has lots of retweets, social interaction, etc. Hence, tweet preprocessing is essential in order to eliminate noisy data that are not useful to the task. We performed some basic text preprocessing listed below:

- 1) Remove metadata: time and ID.
- 2) Remove usernames, mentions, and RTT.
- 3) Remove all numbers.
- 4) Remove redundant texts.
- 5) Remove punctuation.
- 6) Remove foreign characters.
- 7) Remove emojis.
- 8) Remove hashtags.
- 9) Remove repeated characters.
- 10) Remove diacritics (shaddah, fatha, tanwin, damma, kasra, and sukoon)
- 11) Remove tweets that contain URLs or images as their existence could be required to identify any figurative language present in the texts.
- 12) Replace emoticon with its corresponding meaning.
- 13) Use NLTK stopwords
- 14) Normalize some Arabic letters:

- ا إلى ا
- ي to ي
- ء to ئ و
- ك to ك
- ه to ه

### D. Data Annotation

To deal with the problem of [28] that tweets include hashtags are biased and noisy, we manually labelled Both datasets by two Arabic speakers. However, it is considered a time-consuming process. Three main guidelines were given to the annotators:

1- The tweet is ironic if the literal word is opposite to the intended.

2-The tweet is ironic if it was written in a context other than the common context of communication with the aim of negatively mocking sayings, ideas, beliefs or objects.

3- Otherwise, the text is not ironic.

After including the tweets that both annotators agreed on as ironic and non-ironic, the total amount of data are 5620 ironic and 5620 non ironic tweets (Table I). This qualitative analysis revealed that despite having irony-related hashtags, many tweets turned to be not ironic, which shows the significance of manual corpus annotations.

TABLE I. DATASET SUMMARY

| Label        | Number of Tweets |
|--------------|------------------|
| Ironic       | 5620             |
| Non-ironic   | 5620             |
| Total: 11240 |                  |

### E. Features Extraction

We adopted a simple feature extraction technique, which is (pre-trained) word embedding. Word embedding is a form of terms representation for text analysis in which When two words have the same meaning, they are represented in a vector space by similar vectors that are near together. Otherwise, if the terms have different meanings, then the real-valued vectors are far from each other.

To construct such an embedding, Word2Vec [29] is one of the popular techniques. There exist two different methods to learn the embedding: Skip Gram and Continuous Bag of Words (CBOW) (Fig. 2). The CBOW model takes in context words as an input and try to predict the target word equivalent to the context. on the contrary, skip gram tries to predict the surrounding words given a target word which is the opposite of what the CBOW model does. For the purpose of this work, we applied the pre-trained Arabic word embedding model AraVec 3.0 [30], which provides various pre-trained Arabic words. It has a total of 12 distinct word embedding models extracted from various Arabic content domains, which are Twitter, World Wide Web (WWW) pages, and Wikipedia. Moreover, we used Twitter SkipGram 300D-embeddings as according to [29], Skip Gram works good with small dataset and is able to represent uncommon terms. For our neural network model, the embedding vectors are utilized to instate the weights of the embedding layer. Then, it is linked with the remainder of the layers in the system.

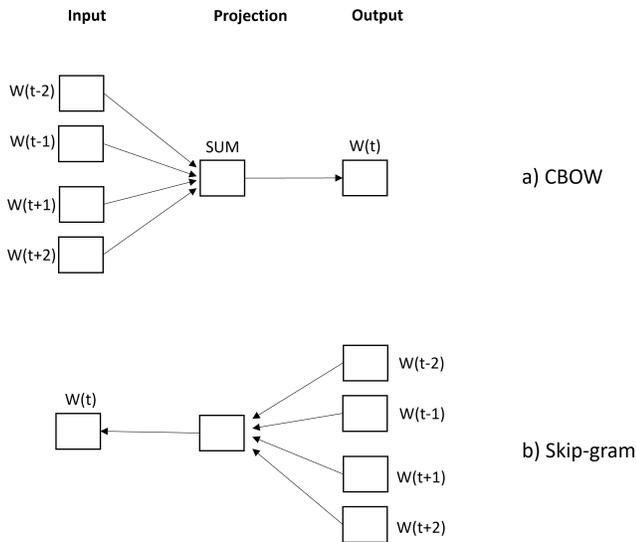


Fig. 2. CBOw and Skip-gram.

F. Methods

1) Convolutional Neural Network (CNN): We adopted a basic CNN architecture, similar to [31], to extract local features from each utterance found in the training dataset by using word vectors that are captured from the pre-trained Word2Vec model. Fig. 3 represents the various layers that are applied to perform the convolution function on the dataset. The Keras library was used to implement the embedding layer. We configured three parameters, which are:

- input\_dim: describes the size of the vocabulary in the data text; it is composed as 30462.
- output\_dim: represents the dimension of the dense embedding; it is configured as 300.
- input\_length: identifies the length of the maximum document; it is composed as 41.

Once we obtained the suitable word vectors from the skip-gram model, the neural network takes the output features as inputs and applies two convolutional layers to the features to learn context information of the words. The convolution operation is described as formula (1).

$$C_i = f(W.X + b) \tag{1}$$

where  $C$  is the convolution output that is created from a window of words  $X$ ,  $W$  is the convolution matrix,  $f$  is the activation function and  $b$  is a bias term.

The two convolution layers have kernel sizes of 4 and 5 to look at sequences of the word embeddings. Each layer consists of 100 filtered outputs. Furthermore, a ReLu activation was applied to the outputs of the convolutional layers followed by a maxpooling layer that takes the highest element from the rectified feature. It minimizes the dimensionality of the feature map and helps capture the key feature. The maxpooling operation (P), as expressed in Equation 3, is done for feature selection that nominates the significant features suitable to

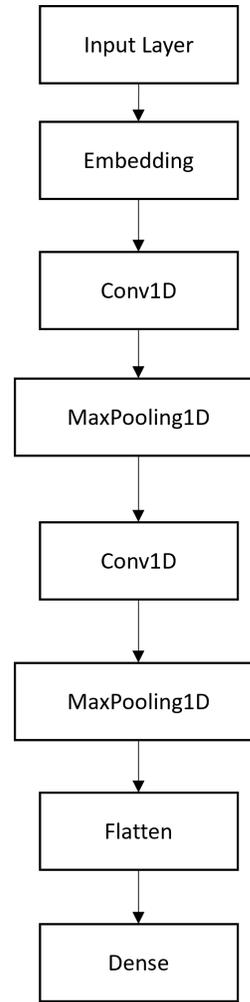


Fig. 3. The Architecture of our CNN.

various hidden layers.

$$C = (c_1, c_2 \dots c_n) \tag{2}$$

$$P_i = \max(C) \tag{3}$$

$$S(x) = \frac{1}{1 + e^{-x}} \tag{4}$$

Next, the output is sent to a layer that flattens the matrices. Finally, the result is a fully connected layer having the outputs of the sigmoid function (Equation 4), that determines whether a sentence is ironic or not, along with the binary cross entropy loss function, which is a good option for binary classification. The model was trained using “Adam” learning rate method.

2) Bidirectional Long Short-Term Memory (BiLSTM): RNN is excellent for sequence learning, but it struggles with long-range dependency due to exploding gradient. On the other hand, LSTM has the capability to learn those long-range dependencies. Bi-LSTM is a variant of LSTM that contains two LSTMs to capture the input information in a forward and backwards directions. In other words, it allows in any point in time to preserve information from both past and future. Given a document,

$$D = (x_1, x_2, \dots, x_n)$$

Bi-LSTM model results with a set of hidden state vectors  $h_t$  for the document sequence. Furthermore, Bi-LSTM combines the forward LSTM (Equation 5) and the backward LSTM (Equation 6).

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \quad (5)$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t+1}) \quad (6)$$

The architecture of our model consists of one Bi-LSTM layer followed by a dropout layer to occupy most of the parameters by ignoring neurons during the training of particular set of units, which is chosen randomly. Next, the model flattens the input data into a 1-dimensional array for inputting to the dense layers. Lastly, two fully connected layers were applied where the first layer contains 32 neurons and the second fully connected layer, which also represents the output layer, contains 2 neurons with the ReLU and Sigmoid activation functions used respectively.

### G. Hyperparameters Setting

There are various important hyperparameters in our systems, and we tune their values using the development corpus. The tuned values of our models hyperparameters are summarized in Table II and Table III.

TABLE II. TUNED VALUES OF THE CNN HYPERPARAMETERS

| Hyperparameter                 | Value               |
|--------------------------------|---------------------|
| Kernel size                    | 4 and 5             |
| Number of filters              | 100                 |
| Maxpooling size                | 2                   |
| Learning rate (adam optimizer) | 0.0001              |
| Loss function                  | Binary crossentropy |
| Batch size                     | 100                 |

TABLE III. TUNED VALUES OF THE BiLSTM HYPERPARAMETERS

| Hyperparameter          | Value               |
|-------------------------|---------------------|
| BiLSTM number of layers | 1                   |
| Number of hidden units  | 128                 |
| Dropout rate            | 0.2                 |
| Optimizer (adam)        | 0.0001              |
| Loss function           | Binary crossentropy |

Using the proper learning rate and setting the correct learning value is critical for enhancing the weights and offsets of the neural model. The low value may cause long training time while its high value could lead to network instability. After several experiments, we set the learning value to 0.0001.

Batch size is another key hyperparameter in neural networks. It defines the number of training examples a neural network can process before resetting the model internal parameters. If the value of batch size too low, then it could slow down

the training process. On the contrary, if the value is too high, it may needs more memory and decrease the generalization capability. Therefore, we started with a small batch value and then increased the value to 64 and 100 in order to use less memory and achieve a durable system.

Neural networks model has the ability of learning complex connections between their inputs and outputs. Yet, some of these relations could be affected by the sampling noise. Hence, the connections will be shown during the training phase but will not occur in the real test data. This problem may cause overfitting and that decreases the classifier's accurate prediction and lead to a poor performance. Therefore, we applied two methods for the purpose of avoiding overfitting in our proposed model. First, we employed early stopping function which refer to stopping the training iterations before the learner passes the point where the model's capability to generalize can decrease as it starts to over-train, thus overfit the training data. Second, we used dropout which is a regularization method of neural networks developed by [32]. It refers to randomly ignore units along with their relationships from the neural network during training process. This stops units from co-adapting too much.

## IV. RESULTS

In this section, we explore the performances of our deep learning models trained with word2vec (skipgram). Table IV shows the performance results of our proposed models. We used the accuracy, recall, precision, and F1 score as performance metrics. These scores are defined as follows [33]:

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (9)$$

$$Accuracy = \frac{TP + TF}{TP + TF + FP + FN} \quad (10)$$

Where TP referred to True Positives, FP is the number of False Positives, FN is False Negatives, and TN is True Negatives. The predicted values are described as positive and negative, while the real values are described as true and false. Recall measure how much out of all the positives were predicted correctly, whereas precision measures how many are actually positive. F-score helps in measuring both recall and precision at the same time. Finally, accuracy measures the overall classifier correctness.

At first, we tried to apply one CNN layer which resulted with 0.86 F1 score. However, the two-layer CNNs slightly improved the performance of the classifier and achieved F1 score of 0.87. We also experimented with BiLSTM and reached 0.86 F1 score, while the results achieved by [12] using BiLSTM model is 0.83. Our scores outperformed the classical and neural approaches used in irony detection in the Arabic language shared task [34]. In spite of the fact that performance of various work existed vary since different datasets have

TABLE IV. RESULTS OF OUR CNN MODELS AGAINST OTHER MODELS

| Model           | Accuracy | Precision | Recall | F1          |
|-----------------|----------|-----------|--------|-------------|
| CNN- One Layer  | 0.86     | 0.90      | 0.83   | 0.86        |
| CNN- Two Layers | 0.87     | 0.90      | 0.84   | <b>0.87</b> |
| BiLSTM          | 0.87     | 0.88      | 0.85   | 0.86        |
| BiLSTM [12]     | -        | -         | -      | 0.83        |

different data distributions, our experiment suggests that the manual annotation improves the learning and prediction tasks. Also, fine tuning the pre-trained vectors and the model play a role in improving the overall results.

Based on the literature [14] [22] [17] [24] [25], CNN's are useful at finding local and position-invariant features whereas BiLSTM are good when classification is specified by a long range semantic relationships and dependency. In our experiment, CNN worked better than BiLSTM in detecting irony since such a sentiment is commonly determined by some key phrases. The output of each convolution layer will let off when a pattern is detected regardless of their position. Changing the size of the kernels and concatenating the received outputs allowed to discover patterns of multiples sizes (4 and 5).

## V. CONCLUSION AND FUTURE WORK

Irony detection research has grown remarkably in recent years. In this paper, we presented an approach based on pre-trained word embedding called AraVec, to address the problem of detecting irony in Arabic tweets. We adopted a basic CNN architecture which found to be very effective for irony detection in Arabic language. The experiment reveals that the method we used performs well with two convolutional layers and even outperform an existing technique. The CNN model achieved F1 score of 0.87. We also experimented with BiLSTM and the results reached 0.86 F1 score. For future work, we plan to extend the extraction of meaningful features, such as sentiment and contextual clues, in order to find the optimal features. Additionally, we can experiment with combined CNNs and RNNs to gain the characteristics of both methods for enhancing the classification performance.

## REFERENCES

[1] K. S. Sabra, R. Zantout, M. A. E. Abed, and L. Hamandi, "Sentiment analysis: Arabic sentiment lexicons," *Sensors Networks Smart and Emerging Technologies (SENSET)*, pp. 1–4, 2017.

[2] P. Rosso, F. Rangel, I. H. Farías, L. Cagnina, and W. Zaghouni, "A survey on author profiling, deception, and irony detection for the arabic language," *wiley*, 2018.

[3] J. Lucariello, "Situational irony: A concept of events gone awry," *Journal of Experimental Psychology*, vol. 2, pp. 129–145, 1994.

[4] R. Filik, A. Turcan, C. RalphNearman, and A. Pitiot, "What is the difference between irony and sarcasm? an fmri study," *cortex*, vol. 115, pp. 112–122, 2019.

[5] A. Bowes and A. Katz, "When sarcasm stings," *Discourse Processes*, p. 215 — 236, 2011.

[6] C. Lee and A. Katz, "The differential role of ridicule in sarcasm and irony," *Metaphor and Symbol*, vol. 13, pp. 1–15, 1998.

[7] E. Sulis, D. Farías, P. Rosso, V. Patti, and G. Ruffo, "Figurative messages and affect in twitter: Differences between irony, sarcasm, and not," *Knowl.-based Syst.*, vol. 108, pp. 132–143, 2016.

[8] R. Filik, E. Brightman, C. Gathercole, and H. Leuthold, "The emotional impact of verbal irony: Eye-tracking evidence for a two-stage process," *Journal of Memory and Language*, vol. 93, pp. 193–202, 2017.

[9] N. Banasik-Jemielniak, A. Bosacki, S. Mitrowska, D. Wyrebek, K. Wisiecka, N. Copeland, L. Wieland, L. Popovic, J. Piper, and A. Siemieniuk, "'wonderful! we've just missed the bus.' – parental use of irony and children's irony comprehension." *PLOS ONE*, vol. 15(2), 2020.

[10] J. Karoui, F. Zitoune, and V. Moriceau, "Soukhria:towards an irony detection system for arabic in social media," *3rd International Conference on Arabic Computational Linguistics*, 2017.

[11] H. A. Nayel, W. Medhat, and M. Rashad, "Benha@idat: Improving irony detection in arabic tweets using ensemble approach," *Proceedings of the IDAT@FIRE2019*, 2019.

[12] M. Khalifa and N. Hussein, "'ensemble learning for irony detection in arabic tweets," *Proceedings of the IDAT@FIRE2019*, 2019.

[13] C. Wu, F. Wu, S. Wu, J. Liu, Z. Yuan, and Y. Huang, "Irony detection with densely connected lstm and multi-task learning," *Proceedings of The 12th International Workshop on Semantic Evaluation*, 2018.

[14] R. A. Potamias, G. Siolas, and A. Stafylopatis, "A transformer-based approach to irony and sarcasm detection," *Neural Computing and Applications*, vol. 32, no. 23, p. 17309–17320, 2020.

[15] P. Golazizian, S. Behnam, A. Seyed, A. Zahra, O. Momenzadeh, and R. Fahmi, "Irony detection in persian language: A transfer learning approach using emoji prediction," *LREC*, 2020.

[16] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," *Proceedings conference on empirical methods in naturallanguage processing*, p. 2539–2544, 2015.

[17] A. Ghosh and T. Veale, "Fracking sarcasm using neural network," *NAACL-HLT*, 2016.

[18] S. Ilic, E. Marrese-Taylor, J. A. Balazs, and Y. Matsuo, "Deep contextualized word representations for detecting sarcasm and irony," *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2018.

[19] E. Riloff, A. Qadir, P. Surve, L. DeSilva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," *EMNLP 2013 conference on empirical methods in natural language processing*, p. 704–714, 2013.

[20] S. Zhang, X. Zhang, J. Chan, and P. Rosso, "Irony detection via sentiment-based transfer learning," *Inform. Proc. and Manag.*, vol. 56, p. 1633–1644, 2019.

[21] J. Gonzalez, L. Hurtado, and F. Pla, "Transformer based contextualization of pre-trained word embeddings for irony detection in twitter," *Information Processing and Management*, vol. 57, 2020.

[22] Y. Huang, H. Huang, and H. Chen, "Irony detection with attentive recurrent neural networks," *Advances in Information Retrieval, Springer*, p. 534–540, 2017.

[23] Y. Ren, D. Ji, and H. Ren, "Context-augmented convolutional neural networks for twitter sarcasm detection," *Neurocomputing*, vol. 308, pp. 1–7, 2018.

[24] A. Baruah, K. B. Das, and K. F. nd Dey, "Context-aware sarcasm detection using bert," *Proceeding 2nd Workshop Figurative Language Process*, pp. 83–87, 2020.

[25] M. S. Razali, A. Halin, L. Ye, S. Doraisamy, and N. M. Norowi, "Sarcasm detection using deep learning with contextual features," *IEEE*, vol. 9, pp. 68 609–68 618, 2021.

[26] TweetScraper, "Jonbakerfish/tweetscraper: Tweetscraper is a simple crawler/spider for twitter search without using api," 2020. [Online]. Available: <https://github.com/jonbakerfish/TweetScraper>

[27] M. Saad, "Arabic sentiment twitter corpus," 2019. [Online]. Available: [https://www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus/data?select=test\\_Arabic\\_tweets\\_positive\\_20190413.tsv](https://www.kaggle.com/mksaad/arabic-sentiment-twitter-corpus/data?select=test_Arabic_tweets_positive_20190413.tsv)

[28] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcastic sentences in twitter and amazon," *In Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pp. 107–116, 2010.

[29] T. Mikolov, S. I., K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, pp. 3111–3119, 2013.

- [30] A. Soliman, K. Eissa, and S. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.
- [31] Y. kim, "Convolutional neural networks for sentence classification," *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.
- [32] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Machine Learning Research*, vol. 15, pp. 1958–7929, 2014.
- [33] M. Hossin and M. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining and Knowledge Management Process (IJDKP)*, vol. 5, 2015.
- [34] B. Ghanem, J. Karoui, F. Benamara, V. Moriceau, and P. Rosso, "Overview of the track on irony detection in arabic tweets," *the 11th Forum for Information Retrieval Evaluation, FIRE '19*, pp. 10–13, 2019.

# Analysis about Benefits of Software-Defined Wide Area Network: A New Alternative for WAN Connectivity

Catherine Janiré Mena Diaz, Laberiano Andrade-Arenas, Javier Gustavo Utrilla Arellano, Miguel Angel Cano Lengua  
Facultad de Ingeniería de Sistemas e Informática  
Universidad Tecnológica del Perú  
Lima, Perú

**Abstract**—This article is based on conducting research to analyze the benefits of emerging trends in communications and networking technology, such as software-defined wide area networks. Using Waterfall as a methodology, the main objective is to carry out a technical comparison at the design and configuration level, creating a virtual environment that simulates traditional and SDWAN (Software-Defined Wide Area Network) infrastructures. The results obtained verify that the benefits of SDWAN maintain business continuity, anticipate situations in which the infrastructure can act intelligently, optimize connectivity while maintaining security, and provide improvements in the management of the entire infrastructure. People will be able to see the results obtained between both technologies and validate the benefits that SDWAN offers.

**Keywords**—Networking technology; connectivity; SDWAN; wide area network; waterfall

## I. INTRODUCTION

Nowadays, for many companies, communications are a vital part of the business continuity, however, in this process the infrastructure and technology that carry out this activity goes unnoticed.

All companies that provide services need to maintain an agile data transmission, as well as an infrastructure that supports it. In turn, connectivity between the company's headquarter, branch offices and data center is a very important aspect because it is necessary to provide services to its customers continuously and with high availability. For this reason, most companies opt to use private networks that are deployed and managed by external service providers, however, today a new technology is being used called SDWAN to communicate all the company's sites. This technology provides a significant change on networking because it delivers innovation in the way it operates, adds a return of investment in medium term and several benefits that makes more companies dare to invest in it [1].

This research work aims to present SDWAN as a new technological alternative for WAN connectivity between multiple sites, because traditionally for 15 years it has been using traditional WAN or known as Multiprotocol Label Switching (MPLS) in Peru. MPLS is known in the technological field as the most popular communication protocol and used by service providers to connect multiple locations of their customers. SDWAN is a relatively new technology in the world since it

is only about 3 years since it became more widely known and it is being adopted by more companies every year.

It is necessary to emphasize that in Peru there are many companies that do not dare to make technological changes or renovations due to a lack of knowledge or rejection of change. Therefore, this research focuses on the benefits that SDWAN technology can bring, for this purpose, the operational benefits at the deployment level of software-defined network solutions for WAN connectivity are analyzed. In addition to the technical benefits at the scalability level and the economic benefits of the software-defined networking solution. Likewise, as support for the research, laboratory tests are carried out that provide technical results in which the benefits that this technology brings in comparison with traditional WAN technology are manifested, not only at a theoretical level, but also at practical level.

This article has the objective to analyze the benefits at the operational, technical, economic level and the comparison between the traditional technology and SDWAN.

## II. LITERATURE REVIEW

According to the growing demand of users towards companies for the use of their services; the use of the internet and the need to maintain connectivity have been fundamental factors in order to deploy applications and meet the expectations of its customers. It can be said that any company that provides services needs to maintain an agile data transmission, as well as an infrastructure that supports it. Nowadays companies are not only looking for availability, reliability and performance so that they can provide connectivity between their sites. Now they are also looking for scalability [2]; for that is inevitable to speak of wide area networks (WAN) when it comes to transmitting data or information. In turn, connectivity, between the sites that a company may have, is a very important issue because it is necessary to provide services to its customers continuously and with high availability. Most companies choose to use private networks that are deployed and managed by external service providers. However, today a new technology is being used to communicate the headquarters of a company, called software-defined wide area networks. This technology provides a change in the use of traditional networks because it delivers innovation in the way it operates, adds a share of cost and benefit in the medium term that makes more companies dare to invest in it

[3]. Finally, it removes the prominence of service providers since it provides greater autonomy to clients to be able to manage their networks [1].

One characteristic that SDWAN has is that it uses three planes: the data plane, which contributes to making communication between sites easier, since a logical infrastructure is created which will work on a physical infrastructure and will allow to transfer data from origin to destinations in a fluid way; the control plane will carry out the control of the configurations of the devices connected to the network, such as policies, routing information, accesses [1]; the orchestration plan that provides a business policy and security framework.

One of the advantages in the use of SDWAN in wide area networks, is that it allows the configuration and management of connections between branches of a company is simple and also flexible, easy to control and supervise, which in the medium and long term results in reducing operating costs [4]. This type of network emerge as a solution to face the different deficiencies that traditional networks present [5], because they facilitate bandwidth management and prioritize data traffic in WAN networks that can also be done in traditional networks, but which requires more effort.

Another advantage that SDWAN networks have is the management, control and configuration of the company's networks from a centralized web platform. This allows changes that are applied to different network connections to take effect immediately. In the centralized platform it will be possible to visualize in a unified way parameters such as the data consumption of the interfaces, latencies of the links, use of the bandwidth by IP, the available bandwidth and other statistics that allow optimal monitoring and control of the network status [4]. It also mentions that in traditional wide area networks to be able to update a network successfully, it is usually necessary to configure the devices manually, which is time consuming and prone to errors compared to an SDWAN type network, in which everything is executed from the centralized platform. Likewise, the platform makes it possible to differentiate the types of traffic that are used by clients or by the companies that manage this platform. These types of traffic are: video traffic, voice traffic, data traffic, and management traffic. Each of them has a prioritization over the other which is defined as follows: video over voice, voice over data and data on administration.

Thus, in a comparison of scenarios to measure and verify the effectiveness of software-defined networks, a simulation was carried out, for which they created a virtual scenario and implemented an SDWAN network. Which allows to establish communication between 2 data centers that use software-defined network technology and in whose tests voice information was transmitted over IP. As a result of the tests carried out, it was evidenced that by configuring the traffic prioritization policies, it is possible to guarantee that the bandwidth quality of service works at an adequate level, in addition to only using a low percentage of CPU load, which translates into efficient network management [5].

In another investigation, 2 scenarios are compared. The first corresponds to a network configured with classic IP / MPLS using manual routing policies and the second scenario uses a software-defined WAN network, for both scenarios the same number of routers and links are used. In the first scenario,

manual routing policies were configured in which low or high latency routes were established. When a router needs to establish communication with another and requests a low latency, the communication will be sent on those links configured with low latency. If at any time the latency of these links increases, the communication will be sent through the same route, as the configuration establishes it. While, in the second scenario, in which a software-defined WAN network is used, the network controller is the one in charge of continuously measuring the status of the links and dynamically defining which route should be used to send the communication [6], as it is programmed in such a way that it has a complete view of the network topology.

In another investigation similar to the one mentioned in the previous paragraph, they evaluated the performance of implementing a software-defined WAN network in an enterprise. In this investigation, they communicated two branches to their headquarters through two Internet / Broadband connections. In the simulation, 2 performance metrics are considered, which are the service time and the percentage of lost packets. The configuration established as the delay time is 10ms and a loss of 2 packets per second (pps). During the simulation, an increase in the delay time was induced, to cause packet losses. When the SDWAN controller detects that the configuration parameters have been exceeded, it automatically establishes a route change for sending the packets, so that it does not affect the quality of service (QoS) [7]. Another investigation refers to the traffic routing associated with applications and administration and how these can be optimized by means of a module added to the controller [8].

Another point to consider in the use of SDWAN technology is that it can work with several connection links, through which traffic is routed through a WAN network making use of load balancing [9], which allows to improve and manage the traffic of the network in the different links [10].

On the other hand, examples of the use of SDWAN in the banking sector are presented. In 2017, there was a drop of 8,800 ATMs in India. Because the satellite connections had a disconnection due to a failure in the satellite leaving many ATMs unusable. This resulted in customers being unable to make transfers or transactions until the failure that cost approximately 600 to 900 billion was resolved. As a lesson learned, it was decided to bet on making viable a project in which important points are considered such as having a connection in contingency in case the main link falls and that is cost effective for the banks. In such a way that it adapts to the existing environment at a technical and cost level, as well as the investment and expected performance, appropriate to the services and the architecture [11]. For this reason, a plan was deployed in the cities of Java and Bali (Indonesia) that in a period of 3.5 years provided satisfactory results and in that same period of time many other companies began to invest in wide area networks defined by software.

The example above explains how SDWAN operates in this specific scenario. Provide the feasibility of having a connection as a contingency in case the satellite connection has any failure. The redundant connection can work in an active / active or active / available way, in either case the objective is to keep the WAN connection secure. The secondary or contingency medium that was chosen in this scenario was 4G / LTE. For

this reason, SDWAN can benefit companies that have many locations and that need to keep their WAN connection always active, in addition, with the TCAC it is demonstrated in an agile way that indicates feasibility and that it can be cost effective.

It is also important to mention that software-defined wide area networks are a relatively new technology and for that reason many companies have not yet chosen to use it. For example, in India, 75% of business customers want to start using these kinds of new solutions; however, only 5% of customers risk making the change, even though it has been shown that this type of network generates great benefits [12], which allows companies to have lower operating costs, higher performance and a robust deployment of software-defined wide area networks. With this technology it can be said that the bottlenecks to transfer information will no longer exist.

Additionally, when comparing traditional WAN networks with SDWAN, the latter are notably superior because in addition to providing centralized control, they allow defining network policies based on profiles and managing data traffic without the need for individual configurations [4]. Also, by massively coupling similar configuration profiles to a set of computers, which would help facilitate the tasks of network administrators because they could manage tasks jointly making more efficient use of work times.

Finally, it is necessary to take into account that security is another outstanding feature in SDWAN because it is present during connectivity, traffic management, additional security services, deployment, visibility and compliance with what is configured [13]. In addition, security is necessary for any company and from the point of view of service providers, security is an essential requirement to be able to implement their projects, although there are different architectures, designs and brands [14].

Based on what was read, it concludes that software-defined WAN networks represent a viable solution to address the difficulties that arise in a traditional WAN network. In addition, it allows to ensure the quality of service through efficient management of network resources, which allows improving network performance, availability and security. Another point to highlight is the reduction of costs in infrastructure and human resources, especially since this technology is simple to administer, which allows a staff with basic knowledge of networks to create offices and perform agile configurations, which shows us that the difficulty in managing networks is significantly reduced, making this type of technology become the next generation network.

### III. METHODOLOGY

The methodology used in this research is Waterfall (see Fig. 1) because the research activities are distributed in phases which are developed sequentially. It is important to mention that a phase cannot be started if the previous one has not been completed. That is why the Scrum framework is not used [15] [16] [17] [18].

#### A. Phase 1 - Start

1) *Project Scope*: Functional requirements: It is necessary to keep in mind that the focus of this research is associated with

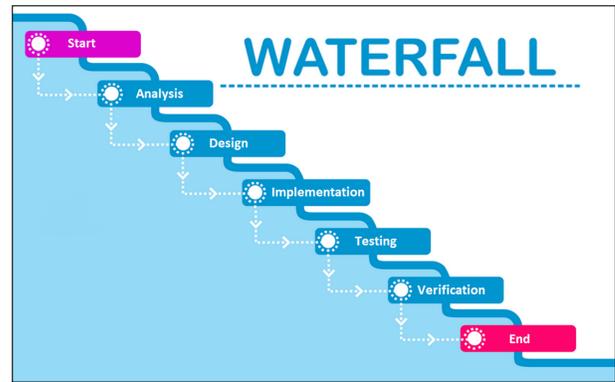


Fig. 1. Implementation Methodology.

the quantitative part, therefore, a laboratory will be executed that will provide results associated with parameters, for this it is necessary to cover certain technical requirements at the software and hardware level to be able to carry out the laboratory.

To perform the deployment for both the traditional environment and the SDWAN environment, PNETLAB tool will be used. This software is a virtual machine that can be provisioned on a VMWare Workstation. The advantage of this technology is that it is not necessary to invest to buy physical devices to be able to emulate the hardware and software.

The laboratory is going to use virtualized devices that are associated with one of the best-known brands in the network and communications environment. Cisco was the chosen brand because its software can be executed in a virtualized environment and also its entire platform is supported and maintained, thus providing reliability for laboratory development. Table I and Table II specify hardware requirements for traditional WAN and SDWAN environments. The hardware and software requirements for the server are detailed in Tables III and IV.

TABLE I. HARDWARE REQUIREMENTS SDWAN

| Device   | Quantity | CPU | Memory | Total CPU | Total Memory |
|----------|----------|-----|--------|-----------|--------------|
| vBond    | 1        | 1   | 1      | 1         | 1            |
| vManage  | 1        | 4   | 16     | 4         | 16           |
| vSmart   | 3        | 2   | 2      | 6         | 6            |
| vEdge    | 3        | 1   | 1      | 3         | 3            |
| vIOS_L2  | 1        | 1   | 1      | 1         | 1            |
| vIOS     | 8        | 1   | 1      | 8         | 8            |
| Computer | 4        | 2   | 8      | 8         | 24           |
| TOTAL    | 22       |     |        | 31        | 59           |

TABLE II. HARDWARE REQUIREMENTS TRADITIONAL WAN

| Device   | Quantity | CPU | Memory | Total CPU | Total Memory |
|----------|----------|-----|--------|-----------|--------------|
| vIOS     | 7        | 1   | 1      | 7         | 7            |
| vIOS_L2  | 1        | 1   | 1      | 1         | 1            |
| vNXOS    | 2        | 1   | 8      | 2         | 16           |
| Computer | 4        | 2   | 8      | 8         | 32           |
| TOTAL    | 14       |     |        | 18        | 56           |

TABLE III. HARDWARE REQUIREMENTS FOR THE SERVER

| Device | CPU(Core) | CPU(vCore) | Memory(GB) | Storage(SSD/TB) |
|--------|-----------|------------|------------|-----------------|
| Server | 16        | 32         | 128        | 1               |

TABLE IV. SOFTWARE REQUIREMENTS FOR THE SERVER

| App                | Version       | Description                |
|--------------------|---------------|----------------------------|
| Windows            | 10            | Operative System           |
| VMWare Workstation | 16.1.2        | Virtual Machine Hypervisor |
| PNETLAB            | 4.2.9         | Simulator Tool             |
| vIOS               | 15.6(2)T      | Virtual Router             |
| vIOS_L2            | 15.2(4.0.55)E | Virtual Switch Layer 3     |
| vNXOS              | 9.2.2         | Virtual Switch Nexus       |
| vManage            | 20.3.2.1      | Virtual Manager            |
| vSmart             | 20.3.2        | Virtual Route Processor    |
| vBond              | 20.3.2        | Virtual Orchestrator       |
| vEdge              | 20.3.2        | Virtual Edge Router        |

2) Project Deliverables:

- Testing scopes
- Work breakdown structure
- Topological diagrams
- Gantt diagram
- Polls
- Survey results
- Results of internet quotes
- Laboratory results

B. Phase 2 - Analysis

1) Traditional Architecture:

Traditional WAN technology is connection-oriented, meaning that its main purpose is to carry information from a source to a destination. In addition, the main weaknesses of the traditional WAN are its manual configuration, its reactive response to incidents, it is oriented to the use of physical components to function and it is not compatible with other programming tools.

Currently, the traditional WAN architecture can cover the different ways of getting from one point to another either by LTE, MPLS, internet or satellite, however, it does not provide dynamism when connecting a point of origin to destination.

Fig. 2 illustrates communication in a traditional WAN environment. It should be noted that communication in a company has the priority of using an MPLS network, in which the components are configured manually. For LTE, internet and satellite, private communications called VPNs are created, which remain manual settings.

2) SDWAN Architecture: For SDWAN, communication is not only based on communicating a point of origin to a destination, but also provides greater benefits such as the automation of configurations, virtual machines can be used and it is compatible with APIs, for this reason it is programmable. In addition, it uses a graphical interface to be able to visualize the health status and settings in a centralized way. Also, it is not oriented to the connection, but to the business need, for example, when a device has two links and one of them is congested, SDWAN has the ability to use a connection that is less congested.

Some of the strengths of SDWAN is that an additional layer of security can be added for information processing and the

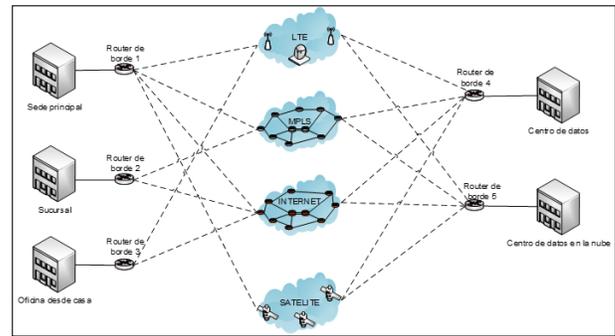


Fig. 2. Traditional WAN Architecture.

provisioning that is given to the equipment is done from a web interface, it is not necessary to configure the equipment onsite, it just requires access to internet and also it needs to be configured with a minimum administration parameters. In addition, it provides analytics on traffic use, applications, users and devices. This makes it a tool that gives companies visibility on bandwidth usage.

Fig. 3 illustrates the communication that occurs between vEdge devices, that are associated with a dynamic communication between any other, also, for each connection a secure and private tunnel is created while the communication is established. The only thing they need is to have ip connectivity between them. The intelligence at the routing level is provided by the vSmart controllers. In addition, you can see the communication between the different premises that, unlike the traditional WAN, is dynamic and provides flexibility when configuring because templates are used and are applied in an automated way.

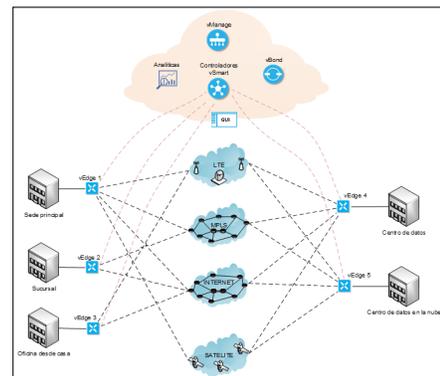


Fig. 3. Cisco SD WAN Architecture.

3) Risks: It is necessary to bear in mind that unexpected events can occur even in a controlled environment, in this case a virtual environment that is simulated by a tool. These risks can be found at the hardware level, software level, or even human error. The most important risks to consider for the development of the laboratory implementation are detailed below.

- Risk 1: Bugs
  - Description: Error caused by SDWAN software code

- Cause: Defective software
  - Consequence: Malfunction in some of the SD-WAN services
  - Probability: Low
  - Impact: High
  - Risk Level: Considerable
  - Control: Use the latest recommended version
- Risk 2: Performance
    - Description: Server saturation for virtual environment emulation
    - Cause: Limited server resources
    - Consequence: Slow use of virtualized computers services
    - Probability: Low
    - Impact: High
    - Risk Level: Considerable
    - Control: Run one virtual environment at a time
  - Risk 3: Human error
    - Description: Error in the configuration of virtualized machines
    - Cause: Disorder in virtual machine configuration
    - Consequence: Lack of connectivity or malfunction in virtualized communication equipment
    - Probability: Medium
    - Impact: Medium
    - Risk Level: Considerable
    - Control: Review the equipment configurations so that a double check is performed

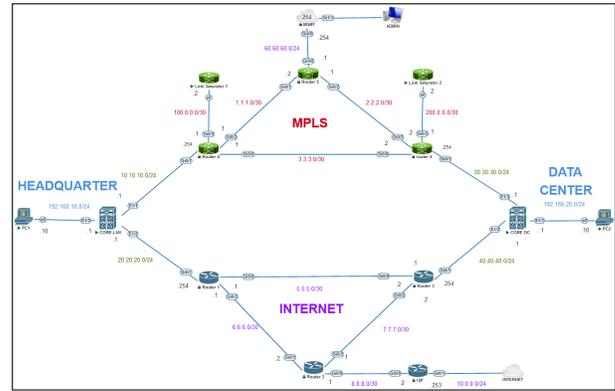


Fig. 4. WAN Topology.

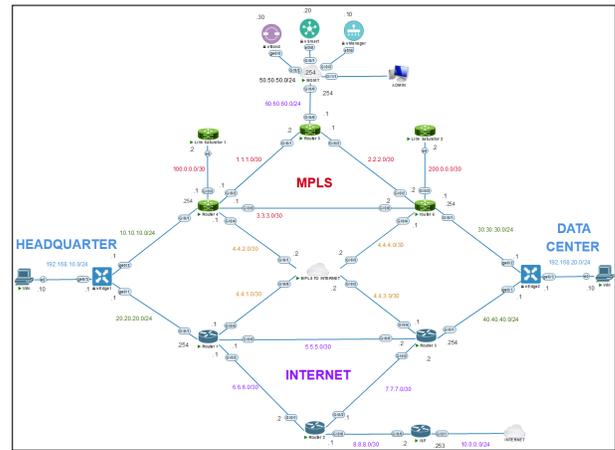


Fig. 5. SDWAN Topology.

### C. Phase 3 - Design

In this phase, a topology is created for traditional WAN and SDWAN environments. Each of the environments will have similar parts as their own technology. It is necessary to keep in mind that the topology for both is logical and physical and also it is oriented for a technical purpose.

A well known design was used in WAN Traditional topology; for SDWAN design MPLS and Internet must have at least a connection between them in order to communicate servers network to Internet interface for vedges.

1) *Technical Scope for Laboratory*: For MPLS and Internet dynamyc routing protocols were used. For practical purpose, OSPF was used.

2) *WAN Traditional Topology* : The connection must be made from Headqueaters to Data center using MPLS and Internet as backup only with MPLS's link is down (see Fig. 4).

3) *SDWAN Topology* : The connection must be made from Headqueaters to Data center using MPLS as a priority and Internet as backup with latency is high (see Fig. 5).

4) *Work Breakdown Structure* : These are the activities used for the laboratory (see Fig. 6).

### D. Phase 4 - Implementation

In the next section shows aspects at the IP addressing level that will be used to run the laboratory and optimize communication between the different communication devices.

1) *Traditional WAN Environment* : The Table V consider the devices shown in Fig. 4.

2) *SDWAN Environment* : The Table VI consider the devics shown in Fig. 5.

## IV. RESULTS AND DISCUSSIONS

### A. Phase 5 - Testing

1) *Testing Traditional WAN*: The tests are based on evaluating two scenarios, one without saturation and the other with saturation in order to evaluate the behavior of a traditional WAN network. The objective of these tests is to validate that the current technology is connection oriented and does not have an intelligence that adapts to changes in the network and can make routing decisions that benefit more stable connectivity in case of intermittences or saturation.

All tests performed consider the path to the MPLS as a routing priority because it is a dedicated network for the client, however, this network in many occasions in a real scenario is

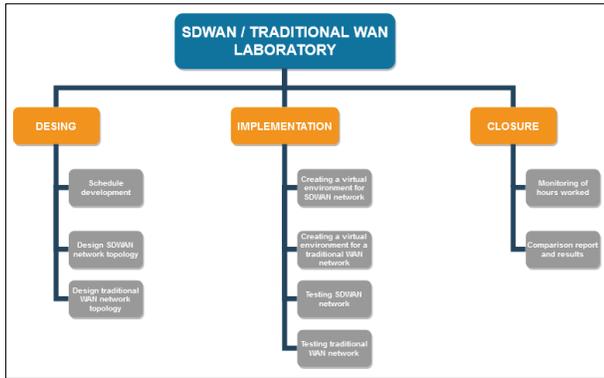


Fig. 6. Work Breakdown Structure.

TABLE V. TRADITIONAL WAN ENVIRONMENT IP ADDRESSING

| HOSTNAME         | LOCATION    | IP/MASK          | GATEWAY      |
|------------------|-------------|------------------|--------------|
| Computer 1       | Headquarter | 192.168.10.10/24 | 192.168.10.1 |
| Computer 2       | Data Center | 192.168.20.10/24 | 192.168.20.1 |
| Link Saturator 1 | Mpls        | 100.0.0.2/30     | 100.0.0.1    |
| Link Saturator 2 | Mpls        | 200.0.0.2/30     | 200.0.0.1    |
| Router 1         | Mpls        | 5.5.5.1/30       |              |
|                  |             | 6.6.6.1/30       |              |
|                  |             | 20.20.20.254/24  |              |
| Router 2         | Mpls        | 6.6.6.2/30       |              |
|                  |             | 7.7.7.1/30       |              |
|                  |             | 8.8.8.1/30       |              |
| Router 3         | Mpls        | 5.5.5.2/30       |              |
|                  |             | 7.7.7.2/30       |              |
|                  |             | 40.40.40.254/24  |              |
| Router 4         | Internet    | 1.1.1.1/30       |              |
|                  |             | 3.3.3.1/30       |              |
|                  |             | 10.10.10.254/24  |              |
|                  |             | 100.0.0.1/30     |              |
| Router 5         | Internet    | 1.1.1.2/30       |              |
|                  |             | 60.60.60.1/24    |              |
|                  |             | 2.2.2.1/30       |              |
| Router 6         | Internet    | 2.2.2.2/30       |              |
|                  |             | 3.3.3.2/30       |              |
|                  |             | 30.30.30.254/24  |              |
|                  |             | 200.0.0.1/30     |              |
| Core LAN         | Headquarter | 10.10.10.1/24    |              |
|                  |             | 20.20.20.1/24    |              |
|                  |             | 192.168.10.1/24  |              |
| Core DC          | Data Center | 30.30.30.1/24    |              |
|                  |             | 40.40.40.1/24    |              |
|                  |             | 192.168.20.1/24  |              |
| ISP              | Internet    | 8.8.8.2/30       |              |
|                  |             | 10.0.0.253/24    | 10.0.0.1     |
| INTERNET         | Internet    | 10.0.0.1/24      | 192.168.1.1  |
| ADMIN            | Internet    | 10.0.0.10/24     | 10.0.0.1     |
| MGMT             | Internet    | 50.50.50.254/24  |              |
|                  |             | 60.60.60.1/30    |              |

also subject to congestion or falls, therefore, a situation will be created in which it is stressed or congested by a lot of traffic.

a) *No link saturation* : Fig. 7 and Fig. 8 shows the Netscantools application after having taken the capture of information for 100 seconds without saturation.

For these tests, ping, latency and jitter and traceability tests have been considered. These tests refer to the hops between devices from a point of origin to a destination. Likewise, Netscantools provides information both at a graphic and textual level, this is very important because it gives visibility of the times that exist while the information goes from one point to another (see Fig. 9).

TABLE VI. SDWAN ENVIRONMENT IP ADDRESSING

| HOSTNAME         | LOCATION    | IP/MASK          | GATEWAY      |
|------------------|-------------|------------------|--------------|
| Computer 1       | Headquarter | 192.168.10.10/24 | 192.168.10.1 |
| Computer 2       | Data Center | 192.168.20.10/24 | 192.168.20.1 |
| Link Saturator 1 | Mpls        | 100.0.0.2/30     | 100.0.0.1    |
| Link Saturator 2 | Mpls        | 200.0.0.2/30     | 200.0.0.1    |
| Router 1         | Mpls        | 5.5.5.1/30       |              |
|                  |             | 6.6.6.1/30       |              |
|                  |             | 20.20.20.254/24  |              |
| Router 2         | Mpls        | 6.6.6.2/30       |              |
|                  |             | 7.7.7.1/30       |              |
|                  |             | 8.8.8.1/30       |              |
| Router 3         | Mpls        | 5.5.5.2/30       |              |
|                  |             | 7.7.7.2/30       |              |
|                  |             | 40.40.40.254/24  |              |
| Router 4         | Internet    | 1.1.1.1/30       |              |
|                  |             | 3.3.3.1/30       |              |
|                  |             | 10.10.10.254/24  |              |
|                  |             | 100.0.0.1/30     |              |
| Router 5         | Internet    | 1.1.1.2/30       |              |
|                  |             | 60.60.60.1/24    |              |
|                  |             | 2.2.2.1/30       |              |
| Router 6         | Internet    | 2.2.2.2/30       |              |
|                  |             | 3.3.3.2/30       |              |
|                  |             | 30.30.30.254/24  |              |
|                  |             | 200.0.0.1/30     |              |
| vEdge1           | Headquarter | 10.10.10.1/24    |              |
|                  |             | 20.20.20.1/24    |              |
| vEdge2           | Data Center | 192.168.10.1/24  |              |
|                  |             | 30.30.30.1/24    |              |
| vManage          | Mpls        | 50.50.50.10/24   | 50.50.50.254 |
| vSmart           | Mpls        | 50.50.50.20/24   | 50.50.50.254 |
| vBond            | Mpls        | 50.50.50.30/24   | 50.50.50.254 |
| ISP              | Internet    | 8.8.8.2/30       |              |
| INTERNET         | Internet    | 10.0.0.253/24    | 10.0.0.1     |
| ADMIN            | Internet    | 10.0.0.1/24      | 192.168.1.1  |
| MGMT             | Internet    | 10.0.0.10/24     | 10.0.0.1     |
|                  |             | 50.50.50.254/24  |              |
|                  |             | 60.60.60.1/30    |              |

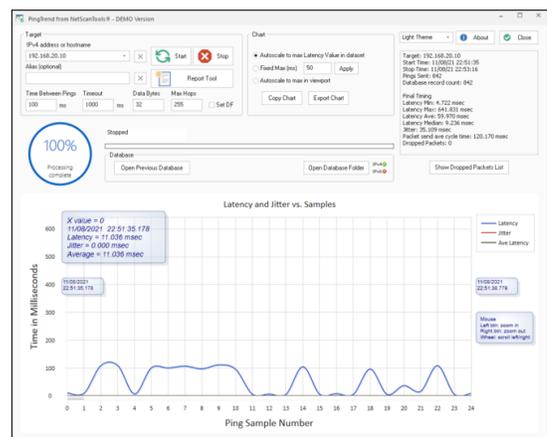


Fig. 7. Test Results without Saturation, no Link Saturation.

b) *With link saturation*: For this part, it was considered an additional program that generates traffic and allows to simulate congestion or saturation in the MPLS network. Fig. 10 show the program that was used to generate traffic on the network.

Fig. 11 and Fig. 12 show the results of the Netscantools application after running the test for 100 seconds with saturation. Fig. 13 shows the last traceroute of the tests.

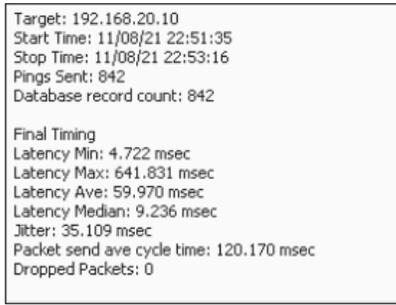


Fig. 8. Closeup of the Results, no Link Saturation.

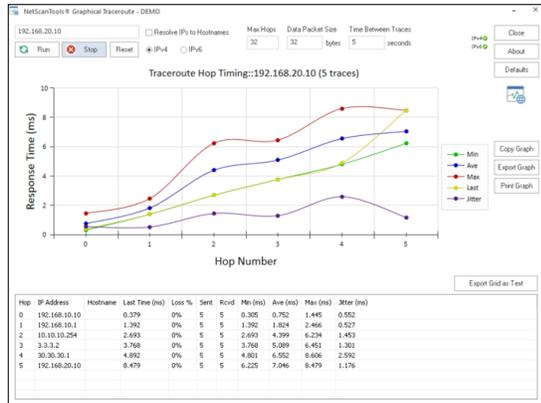


Fig. 9. Five Trace Information.



Fig. 10. Saturation Program Configuration.

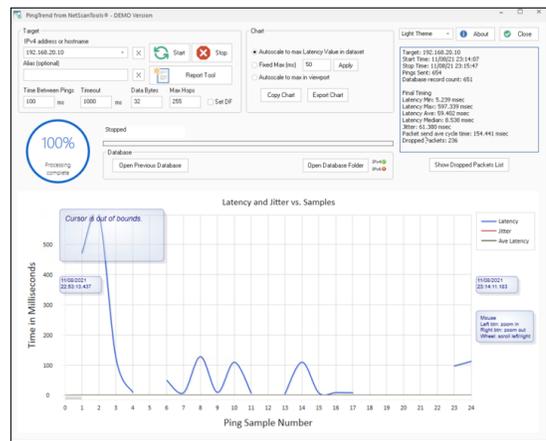


Fig. 11. Test Results of the Netscantools Application with Saturation.

2) *Testing SDWAN*: In this case, two scenarios will also be evaluated, one without saturation and one with saturation in order to find. These tests will validate that the new software-defined network technology allows automatic changes in the network to make routing decisions that benefit more stable connectivity in case of intermittencces or saturation of the links.

a) *No link saturation* : Fig. 14 shows the Netscantools application capture with the resulting information after running the test for 100 seconds without saturation.

Fig. 15 correspond to the graph obtained during the execution time of the test. The values obtained allow to validate that there are minimal changes in the communication speed. It is necessary to take into account that the times are considered in milliseconds.

As in the traditional WAN laboratory, traceability tests have also been considered, which refer to the number of computers through which the information passes from a point of origin to a destination. The tests will be executed in 5 consecutive times, leaving a lapse of 5 seconds between each one. Fig. 16 shows the last traceroute of the tests run.

b) *With link saturation* : Fig. 17 and Fig. 18 are the result after having run the tests for 100 seconds with saturation. Fig. 19 shows the last traceroute of the tests.

3) *Traditional WAN Lab Results* :

a) *Latency*: The following tables compare the results at the latency level obtained in the traces of the tests executed with and without saturation of the links. Table VII compares

latency level with and without saturation. Table VIII and Table IX show the minimum and maximum values obtained in the tests and Table X compares the average values.

TABLE VII. COMPARISON OF THE LEVEL OF LATENCY WITH AND WITHOUT SATURATION WAN

| Hop | IP            | Without Sat. (ms) | With Sat. (ms) | Difference (ms) |
|-----|---------------|-------------------|----------------|-----------------|
| 0   | 192.168.10.10 | 0.379             | 0.281          | -0.10           |
| 1   | 192.168.10.1  | 1.392             | 1.707          | 0.32            |
| 2   | 10.10.10.254  | 2.693             | 790.459        | 787.77          |
| 3   | 3.3.3.2       | 3.768             | 2540.401       | 2536.63         |
| 4   | 30.30.30.1    | 4.892             | 5.298          | 0.41            |
| 5   | 192.168.20.10 | 8.479             | 8.162          | -0.32           |

TABLE VIII. COMPARISON OF MINIMUM WITHOUT AND WITHOUT SATURATION IN MILLISECONDS WAN

| Hop | IP            | Without Sat. Min (ms) | With Sat. Min (ms) | Difference (ms) |
|-----|---------------|-----------------------|--------------------|-----------------|
| 0   | 192.168.10.10 | 0.305                 | 0.281              | -0.02           |
| 1   | 192.168.10.1  | 1.392                 | 1.45               | 0.06            |
| 2   | 10.10.10.254  | 2.693                 | 648.543            | 645.85          |
| 3   | 3.3.3.2       | 3.768                 | 619.221            | 615.45          |
| 4   | 30.30.30.1    | 4.801                 | 5.298              | 0.50            |
| 5   | 192.168.20.10 | 6.225                 | 6.791              | 0.57            |

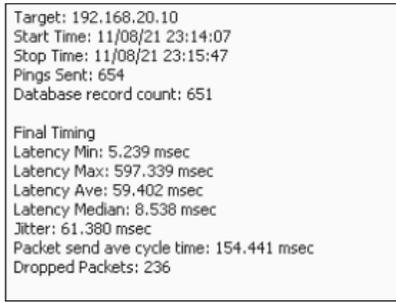


Fig. 12. Closeup of the Results, with Link Saturation.

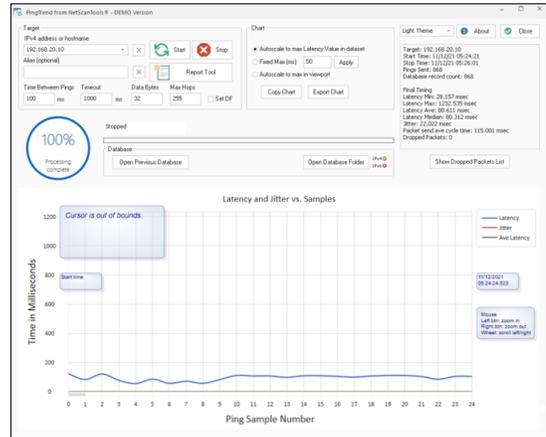


Fig. 14. Netscantools Application Capture with the Resulting Information.

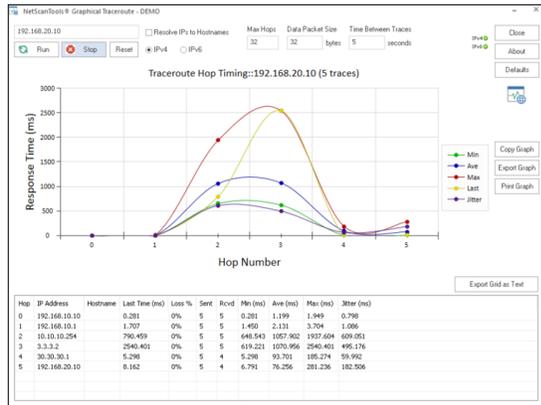


Fig. 13. Fifth Trace Graph of the Test.



Fig. 15. Graph Obtained During the Execution Time of the Test.

TABLE IX. COMPARISON OF MAXIMUM WITHOUT AND WITHOUT SATURATION IN MILLISECONDS WAN

| Hop | IP            | Without Sat. Max (ms) | With Sat. Maxi (ms) | Difference (ms) |
|-----|---------------|-----------------------|---------------------|-----------------|
| 0   | 192.168.10.10 | 1.445                 | 1.949               | 0.50            |
| 1   | 192.168.10.1  | 2.466                 | 3.704               | 1.24            |
| 2   | 10.10.10.254  | 6.234                 | 1937.604            | 1931.37         |
| 3   | 3.3.3.2       | 6.451                 | 2540.401            | 2533.95         |
| 4   | 30.30.30.1    | 6.606                 | 185.274             | 176.67          |
| 5   | 192.168.20.10 | 8.479                 | 281.236             | 272.76          |

TABLE XI. COMPARISON OF JITTER WITHOUT AND WITH SATURATION IN MILLISECONDS WAN ENVIRONMENT

| Hop | IP            | Without Sat Jitter (ms) | With Sat. Jitter(ms) | Difference (ms) |
|-----|---------------|-------------------------|----------------------|-----------------|
| 0   | 192.168.10.10 | 0.552                   | 0.798                | 0.25            |
| 1   | 192.168.10.1  | 0.527                   | 1.086                | 0.56            |
| 2   | 10.10.10.254  | 1.453                   | 609.051              | 607.60          |
| 3   | 3.3.3.2       | 1.301                   | 495.176              | 493.88          |
| 4   | 30.30.30.1    | 2.592                   | 59.992               | 57.40           |
| 5   | 192.168.20.10 | 1.176                   | 182.506              | 181.33          |

b) Jitter: Table XI compares the results of jitter level obtained in the traces of the tests with and without saturation of the links

TABLE X. COMPARISON OF AVERAGE WITHOUT AND WITHOUT SATURATION IN MILLISECONDS WAN

| Hop | IP            | Without Sat. Average (ms) | With Sat. Average (ms) | Difference (ms) |
|-----|---------------|---------------------------|------------------------|-----------------|
| 0   | 192.168.10.10 | 0.752                     | 1.199                  | 0.45            |
| 1   | 192.168.10.1  | 1.824                     | 2.131                  | 0.31            |
| 2   | 10.10.10.254  | 4.399                     | 1057.902               | 1053.50         |
| 3   | 3.3.3.2       | 5.089                     | 1070.956               | 1065.87         |
| 4   | 30.30.30.1    | 6.552                     | 93.701                 | 87.15           |
| 5   | 192.168.20.10 | 7.046                     | 76.256                 | 69.21           |

4) SDWAN Lab Results:

a) Latency: The following tables compare the results at the latency level obtained in the traces of the tests executed with and without saturation of the links. Table XII compares latency level with and without saturation. Table XIII and Table XIV show the minimum and maximum values obtained in the tests and Table XV compares the average values.

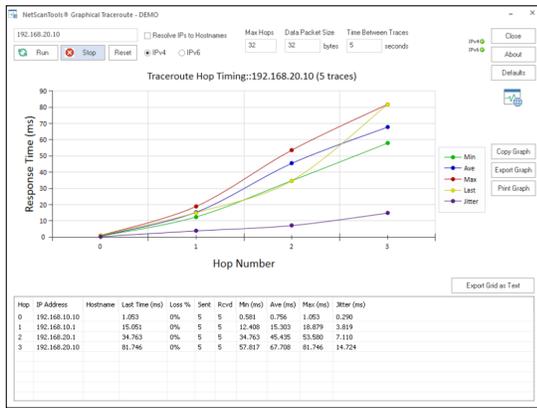


Fig. 16. Last Trace Graph of the Test Run.

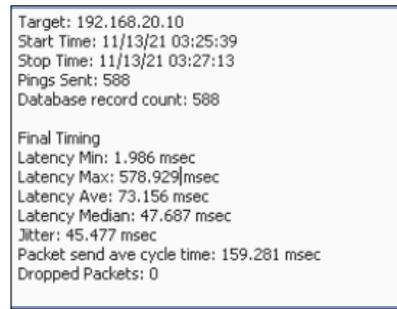


Fig. 18. Ping Test Graph at the Start of the Test.

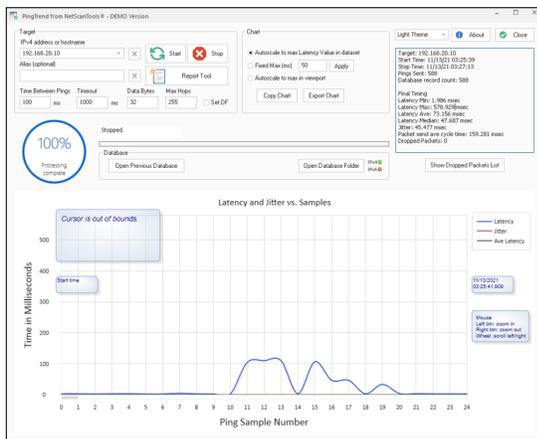


Fig. 17. Test Results with Saturation.

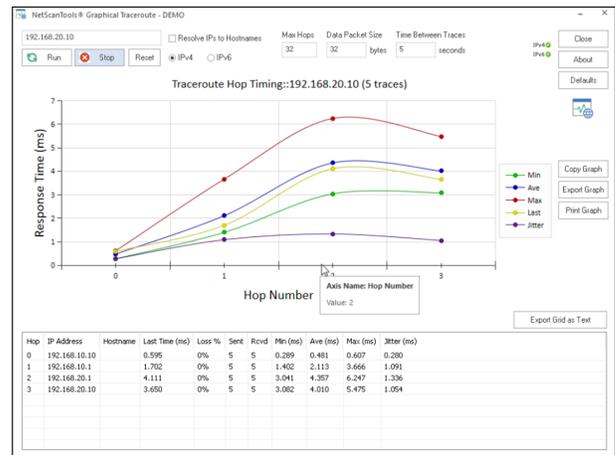


Fig. 19. Last Traceroute of the Test.

TABLE XII. TRACE COMPARISON WITHOUT AND WITH SATURATION IN MILLISECONDS SDWAN

| Hop | IP            | Final Trace (ms) | Final Trace (ms) | Difference (ms) |
|-----|---------------|------------------|------------------|-----------------|
| 0   | 192.168.10.10 | 0.575            | 1.053            | 0.48            |
| 1   | 192.168.10.1  | 1.702            | 15.051           | 13.35           |
| 2   | 192.168.20.1  | 4.111            | 34.763           | 30.65           |
| 3   | 192.168.20.10 | 3.65             | 81.746           | 78.10           |

TABLE XIII. COMPARISON OF MINIMUM WITHOUT AND WITHOUT SATURATION IN MILLISECONDS SDWAN

| Hop | IP            | Without Sat. Minimum (ms) | With Sat. Min (ms) |
|-----|---------------|---------------------------|--------------------|
| 0   | 192.168.10.10 | 0.289                     | 0.581              |
| 1   | 192.168.10.1  | 1.402                     | 12.408             |
| 2   | 192.168.20.1  | 3.041                     | 34.763             |
| 3   | 192.168.20.10 | 3.082                     | 57.817             |

TABLE XIV. COMPARISON OF MAXIMUM WITHOUT AND WITHOUT SATURATION IN MILLISECONDS SDWAN

| Hop | IP            | Without Sat. Maximum (ms) | With Sat. Maximum (ms) | Difference (ms) |
|-----|---------------|---------------------------|------------------------|-----------------|
| 0   | 192.168.10.10 | 0.607                     | 1.053                  | 0.45            |
| 1   | 192.168.10.1  | 3.666                     | 18.879                 | 15.21           |
| 2   | 192.168.20.1  | 6.247                     | 53.58                  | 47.33           |
| 3   | 192.168.20.10 | 5.475                     | 81.746                 | 76.27           |

TABLE XV. COMPARISON OF AVERAGE WITHOUT AND WITHOUT SATURATION IN MILLISECONDS SDWAN

| Hop | IP            | Without Sat. Average (ms) | With Sat. Average (ms) | Difference (ms) |
|-----|---------------|---------------------------|------------------------|-----------------|
| 0   | 192.168.10.10 | 0.481                     | 0.756                  | 0.28            |
| 1   | 192.168.10.1  | 2.113                     | 15.303                 | 13.19           |
| 2   | 192.168.20.1  | 4.57                      | 45.435                 | 40.87           |
| 3   | 192.168.20.10 | 4.01                      | 67.708                 | 63.70           |

b) *Jitter*: Table XVI compares the results of jitter level obtained in the traces of the tests with and without saturation of the links.

TABLE XVI. COMPARISON OF JITTER WITHOUT AND WITH SATURATION IN MILLISECONDS SDWAN ENVIRONMENT

| Hop | IP            | Without Sat Jitter (ms) | With Sat. Jitter(ms) | Difference (ms) |
|-----|---------------|-------------------------|----------------------|-----------------|
| 0   | 192.168.10.10 | 0.28                    | 0.29                 | 0.01            |
| 1   | 192.168.10.1  | 1.091                   | 3.819                | 2.73            |
| 2   | 192.168.20.1  | 1.336                   | 7.11                 | 5.77            |
| 3   | 192.168.20.10 | 1.054                   | 14.724               | 13.67           |

B. Phase 6 - Evaluation

1) *Technical Comparison:* In the next section, the following tables represent the crossing of information between saturation tests of both technologies, also there is a difference in the number of hops, in SDWAN there a less because the connection is by tunneling.

a) *Design:* A big difference between both designs is that SDWAN needs to have servers hosted and Internet and MPLS must have a connection in order to get those servers. In the other hand, for traditional WAN, since each device will be managed independently.

b) *Configuration:* The configuration in the initial phase for SDWAN is executed by command line because it is necessary to assign some parameters for its registration and synchronization with the solution servers. The rest of the configurations are made from a graphical interface. In the other hand, for traditional WAN, 100% of the configurations are made in command lines and additionally the validations are done in the same way. It is important to mention that this can be done initially onsite and then it could be managed remotely.

c) *Latency:* In the results, it can be seen that for traditional WAN technology there is a higher rate of slowness in the response when there is saturation, whereas for SDWAN there is a certain increase in latency, but it is not so considerable (see Table XVII). The important thing is that the connectivity service in general would not be affected and users could perceive a certain slowness, but not disconnections that is a great benefit for SDWAN.

TABLE XVII. LATENCY COMPARISON BETWEEN SDWAN AND TRADITIONAL WAN

|                 | Hop | IP            | Without Saturation | With Saturation  | Difference (ms) |
|-----------------|-----|---------------|--------------------|------------------|-----------------|
|                 |     |               | Final trace (ms)   | Final trace (ms) |                 |
| Traditional WAN | 0   | 192.168.10.10 | 0.379              | 0.281            | -0.10           |
|                 | 1   | 192.168.10.1  | 1.392              | 1.707            | 0.32            |
|                 | 2   | 10.10.10.254  | 2.693              | 790.459          | 787.77          |
|                 | 3   | 3.3.3.2       | 3.768              | 2540.401         | 2536.63         |
|                 | 4   | 30.30.30.1    | 4.892              | 5.298            | 0.41            |
| SDWAN           | 0   | 192.168.10.10 | 8.479              | 8.162            | -0.32           |
|                 | 1   | 192.168.10.1  | 0.575              | 1.053            | 0.48            |
|                 | 2   | 192.168.20.1  | 1.702              | 15.051           | 13.35           |
|                 | 3   | 192.168.20.10 | 4.111              | 34.763           | 30.65           |
|                 | 3   | 192.168.20.10 | 3.65               | 81.746           | 78.10           |

d) *Jitter:* SDWAN has better values because it can switch links in a smart way and this allows the communication to have better performance and quality (see Table XVIII).

TABLE XVIII. JITTER COMPARISON BETWEEN SDWAN AND TRADITIONAL WAN

|                 | Hop | IP            | Without Saturation | With Saturation | Difference (ms) |
|-----------------|-----|---------------|--------------------|-----------------|-----------------|
|                 |     |               | Jitter (ms)        | Jitter (ms)     |                 |
| Traditional WAN | 0   | 192.168.10.10 | 0.552              | 0.798           | 0.25            |
|                 | 1   | 192.168.10.1  | 0.527              | 1.086           | 0.56            |
|                 | 2   | 10.10.10.254  | 1.453              | 609.051         | 607.60          |
|                 | 3   | 3.3.3.2       | 1.301              | 495.176         | 493.88          |
|                 | 4   | 30.30.30.1    | 2.592              | 59.992          | 57.40           |
| SDWAN           | 0   | 192.168.10.10 | 1.176              | 182.506         | 181.33          |
|                 | 1   | 192.168.10.1  | 0.28               | 0.29            | 0.01            |
|                 | 2   | 192.168.20.1  | 1.091              | 3.819           | 2.73            |
|                 | 3   | 192.168.20.10 | 1.336              | 7.11            | 5.77            |
|                 | 3   | 192.168.20.10 | 1.054              | 14.724          | 13.67           |

V. CONCLUSIONS AND FUTURE WORK

SDWAN is a technology that improves capabilities and characteristics over traditional technology in different aspects such as automatization, hardware agnostic, programmable, web interface interaction, business-oriented, cloud-based, secure, scalable and provides analytics. SDWAN networks allow branch offices to have greater communication availability with data centers by prioritizing critical traffic using different types of parameters such as delay or jitter. The result of the tests executed in a virtual environment validates the great benefits at the level of agility and scalability that the solution brings because it was shown that SDWAN was much more effective than the traditional WAN to continue communication despite the fact that the network was congested. It is concluded that SDWAN offers many technical benefits that in the end of the day will keep smart communications and save cost for companies. The next step is to investigate about the return on investment and comparison between the different SDWAN brands that exist in the market so that it can enrich the research presented in this document.

REFERENCES

- [1] S. Badotra, S. N. Panda *et al.*, "A survey on software defined wide area network," *International Journal of Applied Science and Engineering*, vol. 17, no. 1, pp. 59–73, 2020.
- [2] X. Wu, K. Lu, and G. Zhu, "A survey on software-defined wide area networks," *J. Commun.*, vol. 13, no. 5, pp. 253–258, 2018.
- [3] S. Mohmmad, D. Ramesh, S. Pasha, K. Shankar *et al.*, "Research on new network architecture through sd-wan," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 6 Special Issue 4, pp. 483–490, 2019.
- [4] Z. Yang, Y. Cui, B. Li, Y. Liu, and Y. Xu, "Software-defined wide area network (sd-wan): Architecture, advances and opportunities," in *2019 28th International Conference on Computer Communication and Networks (ICCCN)*. IEEE, 2019, pp. 1–9.
- [5] R. E. Mora-Huiracocha, P. L. Gallegos-Segovia, P. E. Vintimilla-Tapia, J. F. Bravo-Torres, E. J. Cedillo-Elias, and V. M. Larios-Rosillo, "Implementation of a sd-wan for the interconnection of two software defined data centers," in *2019 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE, 2019, pp. 1–6.
- [6] I. Šeremet and S. Čaušević, "Advancing ip/impls with software defined network in wide area network," in *2019 International Workshop on Fiber Optics in Access Networks (FOAN)*. IEEE, 2019, pp. 56–61.
- [7] S. Troia, L. M. M. Zorello, A. J. Maralit, and G. Maier, "Sd-wan: an open-source implementation for enterprise networking services," in *2020 22nd International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2020, pp. 1–4.
- [8] H. Zhang, Y. Wang, X. Qi, W. Xu, T. Peng, and S. Liu, "Demo abstract: An intent solver for enabling intent-based sdn," in *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, 2017, pp. 968–969.

- [9] S. Rajagopalan, "An overview of sd-wan load balancing for wan connections," in *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 2020, pp. 1–4.
- [10] P. Segeč, M. Moravčík, J. Uratmová, J. Papán, and O. Yeremenko, "Sd-wan-architecture, functions and benefits," in *2020 18th International Conference on Emerging eLearning Technologies and Applications (IC-ETA)*. IEEE, 2020, pp. 593–599.
- [11] S. Andromeda and D. Gunawan, "Techno-economic analysis from implementing sd-wan with 4g/lte, a case study in xyz company," in *2020 International Seminar on Intelligent Technology and Its Applications (ISITIA)*. IEEE, 2020, pp. 345–351.
- [12] S. Sinha, R. Chowdhury, A. Das, and A. Ghosh, "Prospective sd-wan shift: Newfangled indispensable industry driver," in *Soft Computing Techniques and Applications*. Springer, 2021, pp. 255–261.
- [13] M. Wood, "Top requirements on the sd-wan security checklist," *Network Security*, vol. 2017, no. 7, pp. 9–11, 2017.
- [14] —, "How to make sd-wan secure," *Network Security*, vol. 2017, no. 1, pp. 12–14, 2017.
- [15] A. Carrion-Silva, C. Diaz-Nunez, and L. Andrade-Arenas, "Admission exam web application prototype for blind people at the university of sciences and humanities," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0111246>
- [16] R. Arias-Marreros, K. Nalvarte-Dionisio, and L. Andrade-Arenas, "Design of a mobile application for the learning of people with down syndrome through interactive games," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0111187>
- [17] A. Tupia-Astoray and L. Andrade-Arenas, "Implementation of an e-commerce system for the automation and improvement of commercial management at a business level," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120177>
- [18] V. Gomero-Fanny, A. R. Bengy, and L. Andrade-Arenas, "Prototype of web system for organizations dedicated to e-commerce under the scrum methodology," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120152>

# Data Recovery Approach for Fault-Tolerant IoT Node

Perigisetty Vedavalli, Deepak Ch  
School of Electronics and  
Communication Engineering  
VIT-AP University  
Amaravathi, Andhrapradesh, India 522237

**Abstract**—Internet of Things (IoT) has a wide range of applications in many sectors like industries, health care, homes, militarily, and agriculture. Especially IoT-based safety and critical applications must be more securable and reliable. Such type of applications needs to be operated continuously even in the presence of errors and faults. In safety and critical IoT applications maintaining data reliability and security is the critical task. IoT suffers from node failures due to limited resources and the nature of deployment which results in data loss consequently. This paper proposes a Data Recovery Approach for Fault Tolerant (DRAFT) IoT node algorithm, which is fully distributed, data replication and recovery implemented through redundant local database storage of other nodes in the network. DRAFT ensures high data availability even in the presence of node failures to preserve the data. When an IoT node fails in any cluster in the network data can be retrieved through redundant storage with the help of neighbor nodes in the cluster. The proposed algorithm is simulated for 100-150 IoT nodes which enhances 5% of network lifetime, and throughput. The performance metrics such as Mean Time to Data Loss (MTDL), throughput, Network lifetime, and reliability are computed and results are found to be improved.

**Keywords**—Internet of things; data recovery; RAID; node failures; reliability; network lifetime

## I. INTRODUCTION

IoT provides an integration of multiple controllers, IoT nodes, servers, and gateways which contains embedded technologies to be logically connected and enables them to sense and interact to the real world and also among themselves. It is attainable for them to gather data from a wide range of existing structures. The accuracy of the IoT network is diminished, when the transmission data is faulty which leads to cause for inappropriate actions. So, it is critical to enhancing the ability of the IoT network to detect and recover the faulty node's data. Difficulties of detecting incorrect data and the quality of data have been studied extensively [1]. Fault tolerance is an important aspect for ensuring the high reliability and availability of the IoT network. Due to resource-constrained devices, there may be a lot of chances to occur failures in the network. Hence traditional networking technologies cannot handle IoT requirements effectively [2]. According to the survey, 48% of IoT projects may fail due to data failures and security. Data failures include missing files, corrupt files, and data blocks, and inconsistent files [3].

Capturing the data quality levels is more effective to estimate the device's quality and their produced data. Data quality estimation mechanism depends on three stages[4],

data reliability, device availability, and overall quality of data. Sensors are small devices with limited resources like memory space, battery, and processing power. These resource-constrained nodes pose multiple challenges for network designers' accurate usage of scarce resources. In certain times IoT applications need to be deployed in harsh environments [5]. In such IoT applications, nodes are prone to failures due to various reasons, such as hardware failure, battery depletion, and external events. Complex IoT applications with the help of various techniques require effective data management [6]. IoT is a complex heterogeneous network, maintaining high reliability is one of the major concerns. IoT networks should be more reliable for safety and critical applications, with the help of heuristic binary decision diagrams can able to access the link failures that occur due to data loss in community structures [7].

IoT applications help the industries to bring a competitive edge on their competitors. Even due to sharing data, security, device faults, and data manipulation between the various smart devices which becomes a serious concern to many industries, these interrupt the workflow of industries. IoT network comprises many sensing nodes, so the network needs to collect and process an enormous amount of external environment data. Enhancing the fault tolerance and reliability can be achieved by adding redundant bits to the original data at the information level is necessary for an IoT network. By employing the Reduced Variable Neighborhood Search (RVNS) algorithm the IoT network can enhance the processing speed and reliable transmission of data [8]. Generally, sensor data validation includes mainly two steps, those are data faults detection and data reconciliation. There is no perfect tool or method for this process.

In [9] various faulty data detection and correction methods and tools are discussed. In an IoT network, each sensor node works with a limited power source and when the sensor stops working the network cannot process data that they received this may affect the prediction of network health and reliability which leads to network failure. When this situation happens RAID structures are very useful for maintaining the redundant copies of data. This paper focuses on data management with the help of a particular RAID-like technique in the cluster for achieving fault tolerance with additional communication costs.

The remaining paper is organized as Section 2 recent data replication and recovery methods in IoT, Section 3 is about the implementation of a DRAFT algorithm in the IoT network. Section 4 is discussed about performance evaluation for DRAFT Algorithm and finally, section 5 draws conclusion

and possible future directions of work.

## II. RELATED WORK

IoT requires efficient data collection, generation, and presentation through wireless sensor nodes. Due to limited energy resources, sensor nodes are unreliable, which may lead to the loss of valuable data. In IoT, data replication is a promising method for data management. In [10] data replication algorithms for IoT nodes classification, analysis, and comparison are presented. Data replication techniques include query balancing, data availability, system robustness, and data retrieval Resilient data-centric storage algorithm is utilized by [11] to illustrate the tiny database systems for easy data retrieving. It is specially designed for low-powered IoT sensor nodes. A Distributed Hash Table(DHT) is used for storing the redundant data into the node.

The redundant storage of data assures the information available in case of source failure. A greedy replication-based distributed storage algorithm is proposed in [12], if a node in the network fails then the data can be retrieved through neighbor donor nodes. IoT sensor nodes data is stored in distributed mini data centers in a decentralized manner cloud rather than single cloud storage. In those data storage areas, each IoT data item has predetermined redundant data copies. This problem has been formulated with the help of a complex-linear programming model and heuristic algorithms are proposed in [13]. These algorithms are helping to improve the latency of reading and writing operations.

A multi-dimensional data storage algorithm is implemented within a single node [14], which can create to handle querying, indexing, storing, and ingestion of huge amounts of data. A distributed MDDS offers high ingestion rates, fault tolerance, and horizontal scaling when compared to Relational database management systems (RDBMS). To assure high data availability in the IoT network during the node failure distributed hop by hop data replication (DRAW) technique is proposed in [15]. This algorithm helps to identify the best replica node for maintaining a redundant copy. For selecting the replica node this technique applies a series of conditions like availability of the memory in the device, the number of hops, degree of replication, previous replicas of the data items, and common neighbors of the devices.

Data availability during the failures can be achieved through data replication algorithms. In [16] bridged replication control algorithm (BRC) is proposed for smart logistics. BRC creates temporary replication access when the link failure occurred through the bridge token. BRC provides efficient data management for smart networks. By maintaining redundant data components in multiple storage locations can achieve high fault tolerance. An adaptive data replication algorithm is proposed [17], and incorporated at the gateway level.

IoT network is deployed with many sensor nodes, in this context energy of the sensor nodes is depleted during data transmission. An adjustable data replication algorithm based on virtual grid technology is implemented in [18]. This scheme helps to enhance the lifetime and performance of the sensor nodes. A dynamic sink node will determine the communication link depending on the selected beacon and continuously develop a replica node across the query node

TABLE I. COMPARISON OF VARIOUS FAULT DATA RECOVERY APPROACHES AND REPLICATION TECHNIQUES

| Author             | Method                                                | Single/<br>Multi<br>Node<br>point<br>Failure | Fault<br>recovery<br>and<br>prediction     | Replica<br>Distribution                 | No.of<br>Nodes   |
|--------------------|-------------------------------------------------------|----------------------------------------------|--------------------------------------------|-----------------------------------------|------------------|
| Shong[9]           | Edisense                                              | Multi<br>point<br>failure                    | Fault<br>handling<br>Mechanism             | Locality<br>based                       | 10-20<br>Nodes   |
| Qaim[13]           | Distributed<br>hop by hop<br>replication<br>algorithm | Single<br>point<br>failure                   | Data<br>errors<br>handling                 | Connectivity<br>energy<br>based         | 50-70<br>Nodes   |
| Juan[14]           | Low<br>complexity<br>greedy<br>algorithm              | Single<br>point<br>failure                   | Data<br>errors<br>handling                 | Memory<br>based                         | 10-50<br>Nodes   |
| Qaim[15]           | Adjustable<br>data<br>replication<br>algorithm        | Single<br>point<br>failure                   | Data<br>errors<br>handling                 | Uniform<br>based                        | 100<br>Nodes     |
| DRAFT<br>algorithm | RAID based<br>replication<br>algorithm                | Single<br>point<br>failure                   | Data<br>errors<br>and<br>fault<br>handling | Probabilistic<br>and<br>memory<br>based | 100-150<br>Nodes |

to create a balance between the rate of energy consumption and the overhead in the network. Data recovery is one of the essential features of an IoT network. These networks may face some issues due to sensing and connection errors which result in incorrectly received data [19]. By incorporating probability matrix factorization at the cluster level can recover the missing data through neighbor nodes in the cluster.

In [20] a convolution neural network has been utilized for generating data recovery algorithms. For restoring the data which mainly includes two steps, all sensed collected for the training process to the networks and data recovery has been initiated with the help of a trained generator. An redundant residue number system algorithm [21], Max-flow algorithm fault tolerant [22], Device pairing algorithm [23], Finding Least connected points algorithm [24], Least connected neighbour algorithm [25] are the few fault-tolerant algorithms have been studied. The above-listed literature algorithms in Table I were presented with a minimum of 10 to a maximum of 100 nodes. This work was analyzed properly to extend the existing algorithms works towards 150 IoT nodes with network lifetime, MTTDL, Reliability, and throughput of the network.

## III. IMPLEMENTATION OF A DRAFT ALGORITHM IN THE IOT NETWORK

Safety and critical industrial IoT applications such as IoT based nuclear radiation monitoring systems require more data availability and high fault tolerance during data loss. Thus protecting sensed data through redundant storage can significantly improve the system performance. Let us consider 'n' clusters with 'N' static nodes are randomly deployed in a sensing field. Each node has a unique ID, 'i' where  $i=1,2,3,\dots,P$ . Each node 'i' has fixed memory space 'k', to hold data and parity items. The memory space is partitioned into two units based on the RAID5 structure, one is for storing data and another unit is for parity storage. RAID is a well-proportioned technology that creates improved storage reliability and functions by block-level striping with parity in the node storage area. At

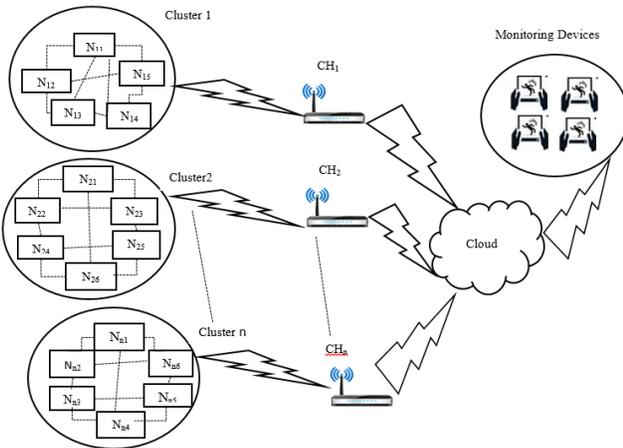


Fig. 1. Cluster Based Multi-Node IoT Architecture.

regular intervals, all the nodes detect environmental factors and produce sensing data  $D_i$ . The effective capacity for storing the sensing information is  $(k-1)/k$  of its total storage 'k'.  $1/k$  storage space is utilized for storing redundant information. Each node preserves a Direct Neighbor Node(DNN) attribute list ' $N_{Att}$ '. Every sensor node in the network can produce data and updates in the data unit as well as parity unit, with the help of all DNN in the cluster, and the node failure is denoted as  $N_F$ . To implement the DRAFT algorithm, a few assumptions have been made regarding the cluster-based multi-node IoT architecture as follows

- Assume that each cluster has at least five nodes.
- All nodes are having identical significance and storage space.
- All the nodes are having similar computation resources and initial energy supply.

To follow these steps the architecture is built as shown in Fig. 1. The basic architecture of IoT nodes consists of five elements they are controller, sensing element, battery, local storage, and network connectivity. The local storage of the node is partitioned into two parts. A DRAFT algorithm is invoked at each node and its storage unit is partitioned into two units based RAID5 mechanism. The inputs of the algorithm are no. of clusters 'n', storage space 'k', Data items  $D_i$ ,  $1/k$  parity unit,  $(k-1)/k$  data unit space, DNN attribute table  $N_{Att}$ . As the outputs of the algorithm create a node ID, and generate the parity based on DNN, data recovery from the DNN, MTTDL, throughput, network lifetime, and reliability. For every round, if any node data loss occurs can be retrieved through the DNN of that particular node in the cluster. A node collects new data samples continuously, and that data only updates its  $D_i$  section, but it also transmits a copy of the newly detected data to all of its direct neighbors, allowing them to update their  $P_i$  sections by computing the parity from the received data. The procedural steps for DRAFT Algorithm in IoT network as follows:

### DRAFT Algorithm

#### Step1 Input declaration

**Input:**Number of clusters n, Data item  $D_i$ , storage space k, DNN attribute table  $N_{Att}$

#### Step2 Output declaration

**Output:**Parity generation, data recovery if any data loss occurs,MTTDL,Network Lifetime, throughput,and reliability.

#### Step3 Node structure

Algorithm defnode (node id, cluster id, Data  $D_i$ , parity  $P_i$ )

begin

$C_{id}$  = Cluster id

$N_{id}$  = Node id

$D_i$  = Data of ith node

$P_i$  = Parity of ith node

end

#### Step4 Storage unit structure

Algorithm DRAFT (n, k, DNN)

def node

Data unit =  $(k-1)/k$

Parity unit =  $1/k$

update  $D_i, N_{att}, P_i$

#### Step5 Initialize the first round

Node i transmitted data to cluster head i

if ( $D_i$  of  $N_i = D_i$  of  $C_i$ )

transmit ACK

{

initialize next round

}

Else (error detection=true)

Transmits NACK

{

For each node in cluster A data recovery

$$DL_{iA} = \sum_{j=1}^n D_{jA}$$

$P_i = \text{XOR}(\text{DNN data of } i^{\text{th}} \text{ node})$

Data Recovered successfully

transmit ACK

}

#### Step6 Network lifetime extraction

Network lifetime

{

$P_v = (1-p)^{N_d}$

}

#### Step7 MTTDL

{

MTTDL =  $\int_0^{\infty} P(S_0(t))+P(S_1(t))+P(S_2(t))dt$

$$MTTDL = \frac{H(x)}{V(x)}$$

Where as

$H(x) = (\mu_D + k\lambda_0 + (k-1)\lambda_1)(\alpha_1 + (k-1)\beta_1) + (k\lambda_0 + (k-1)\lambda_1)(\mu_D + \lambda_2 + (k-1)\lambda_1)$

$V(x) = (k\lambda_0(k-1)\lambda_1(\alpha_1 + \lambda_2 + (\mu_D(k-1)\lambda_1)(k-1)(\lambda_1 + \beta_1))$

if  $\mu_D \rightarrow \infty$

{

$MTTDL = \frac{\alpha_1 + k\lambda_0 + (k-1)(\lambda_1 + \beta_1)}{k\lambda_0(k-1)(\lambda_1 + \beta_1)}$

}

if  $\mu_D \rightarrow 0$

$$MTTDL = \frac{1}{k\lambda_0} + \frac{1}{(k-1)\lambda_1}$$

**Step8 Reliability extraction**

$P(S_0(t)) + P(S_1(t)) + P(S_2(t)) + P(S_F(t)) = 1$   
For each cluster A in N {

$$\frac{\partial P(S_0(t))}{\partial t} = -k\lambda_0 P(S_0(t)) + \alpha_1 P(S_2(t))$$

$$\frac{\partial P(S_1(t))}{\partial t} = k\lambda_0 P(S_0(t)) - (\mu_D + (k-1)\lambda_1)P_1(t) + \lambda_2 P(S_2(t))$$

$$\frac{\partial P(S_2(t))}{\partial t} = \mu_D P(S_1(t)) - (\alpha_1 + \lambda_2 + (k-1)(\lambda_1 + \beta_1))P(S_2(t))$$

$$\frac{\partial P(S_F(t))}{\partial t} = (k-1)\lambda_1 P(S_1(t)) + (k-1)(\lambda_1 + \beta_1)P(S_2(t))$$

**Step9 Throughput**

if  $D_i$  of  $N_i = D_R$   
Successful recovery  
else  
Unrecoverable data  
Repeat  
End

Parity is computed by XORing all the inputs in a bit-wise manner. If there are more than one boolean inputs, XOR returns true when the two inputs are different. A parity scheme is one of the common approaches for error detection. Cluster1 is grouped with five nodes each node is connected with three nodes. If any node failure causes the neighbor nodes will help to retrieve the data of that failed node by using redundant information. Data storage in  $D_i$  and  $P_i$  and recovery is represented in Fig. 2 topology. Here defining the following relations for nodes with their corresponding neighbors,

$$P_1 = D_1 \oplus D_2 \oplus D_3 \oplus D_5 \tag{1}$$

$$P_2 = D_1 \oplus D_4 \oplus D_3 \oplus D_2 \tag{2}$$

$$P_3 = D_1 \oplus D_4 \oplus D_5 \oplus D_3 \tag{3}$$

$$P_4 = D_2 \oplus D_3 \oplus D_5 \oplus D_4 \tag{4}$$

$$P_5 = D_1 \oplus D_4 \oplus D_3 \oplus D_5 \tag{5}$$

By using the above equations 1 to 5, the pi will store the information in cluster 1, similarly, all the clusters in the network the parity unit will update based on the DNN of the particular node in the network. Di unit will update the sensing data into it.

Using parity can easily rebuild the lost data inputs by conducting XOR of all the remaining values and former output values. Assume that node2 fails the data recovery of the node2

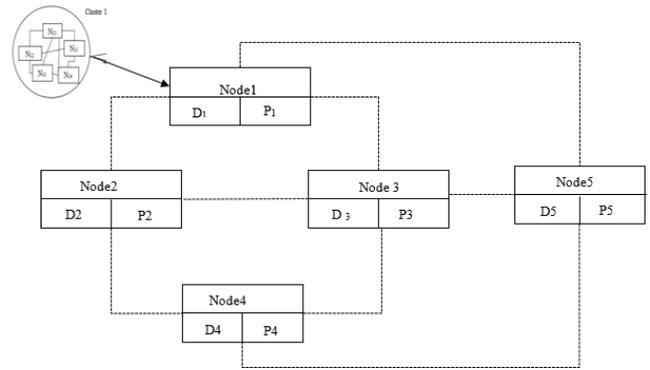


Fig. 2. Representation of Cluster1 Data Storing and Communication Process.

TABLE II. DATA RECOVERY PATTERNS OF CLUSTER1 FOR ROUND1 AND ROUND2

| Cluster1          |          |                |                            |                                             |
|-------------------|----------|----------------|----------------------------|---------------------------------------------|
| Transmission Data | Parity   | Receiving Data | Faulty Data Identification | Data Recovery                               |
| D1 = 100          | P1 = 111 | 100            | ACK                        | No error                                    |
| D2 = 101          | P2 = 101 | 111            | NACK                       | error occurred, Recovery initiated, D2= 101 |
| D3 = 111          | P3 = 001 | 111            | ACK                        | No error                                    |
| D4 = 011          | P4 = 000 | 011            | ACK                        | No error                                    |
| D5 = 001          | P5 = 001 | 001            | ACK                        | No error                                    |
| D1 = 101          | P1 = 100 | 101            | ACK                        | No error                                    |
| D2 = 101          | P2 = 011 | 101            | ACK                        | No error                                    |
| D3 = 111          | P3 = 101 | 111            | ACK                        | No error                                    |
| D4 = 100          | P4 = 101 | 100            | ACK                        | No error                                    |
| D5 = 011          | P5 = 101 | 001            | NACK                       | error occurred, Recovery initiated, D5= 011 |

will get with the help of redundant information stored and its corresponding neighbors, Identified that if D2 and D5 at round1 and round2 instances then data recovery has been done with the help of all DNN and parity of the failed node in that cluster which is represented in Table II. Hence the data is proved that fault recovery has been made using a DRAFT algorithm. IoT network is grouped with 'n' no. of clusters, each cluster consists of  $n \geq 5$  nodes. Each node has four states they are normal state, degraded state, recovery state, and failed state. The reliability of the network and MTTDL of the node has been derived with the help of the following state diagram Fig. 3.

Let us consider the nonzero time of node replacement, the failure rate of a node is normal in degraded and rebuild states.  $S_0$  is the normal state, in this state all the nodes in the cluster are operable and data in every node is available. From this state, the node can pass to state  $S_1$  with the rate of  $k_0$ (Failure of any node).

$S_1$  is a degraded state, in this state one of the nodes in the cluster has been failed and waiting for a replacement and the remaining k-1 nodes are operable and data can be available. From this state, the node can move to either state F with the failure rate  $(k-1)\lambda_0$  (failure of another operable node) or  $S_2$  with the repair rate of  $D$ .

$S_2$  is the recovery state, in this state the failed node is replaced and the recovery process has been started, the remaining k-1

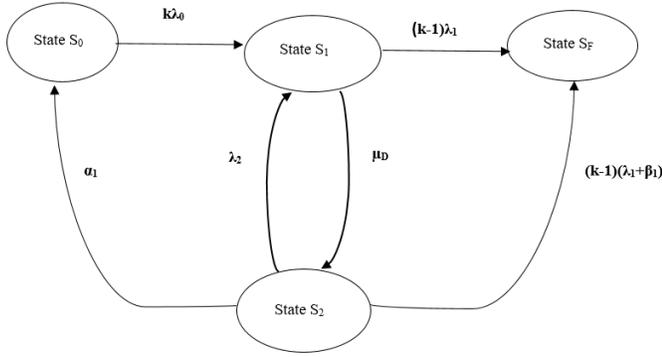


Fig. 3. Reliability State Graph Model for An IoT Cluster.

nodes are operable and data is available. From this state, the node can move to either state0 with the rate of  $\alpha_1$  (successfully data recovery completed) or to  $S_1$  with the failure rate  $\lambda_2$  (failure of the node during the data recovery process) or to  $S_F$  with the rate of  $(k-1)\lambda_1$  (failure of one of the operable nodes) or with the rate of  $(k-1)\beta_1$  (read error on one of the operable nodes during data recovery process).  $S_F$  is the failed state data of the node that are non-recoverable and unavailable. Let us present the state diagram based on the above transitions.

Where as

- $\lambda_0$  is the failure rate of the node in a cluster.
- $\lambda_1$  is the failure rate of the node in case of data unavailability of the node which is operable.
- $\lambda_2$  is the failure rate of the replaced node during the data recovery.
- $\mu_D$  repair rate of the faulty node.
- $\alpha_1$  is the success rate of data recovery.
- $\beta_1$  rate of reading errors on the operable nodes during the data recovery.

The above state diagram is solved with the help of Kolmogorov-Chapman differential equations which are analyzed as follows:

Initially the probabilities of state  $S_0$  is  $P(S_0(t)) = 1$ , the remaining states probability which is equals to zero, hence  $P(S_1(t)) = P(S_2(t)) = P(S_F(t)) = 0$ .

$$P(S_0(t)) + P(S_1(t)) + P(S_2(t)) + P(S_F(t)) = 1 \quad (6)$$

$$\frac{\partial P(S_0(t))}{\partial t} = -k\lambda_0 P(S_0(t)) + \alpha_1 P(S_2(t)) \quad (7)$$

$$\frac{\partial P(S_1(t))}{\partial t} = k\lambda_0 P(S_0(t)) - (\mu_D + (k-1)\lambda_1) P(S_1(t)) + \lambda_2 P(S_2(t)) \quad (8)$$

$$\frac{\partial P(S_2(t))}{\partial t} = \mu_D P(S_1(t)) - (\alpha_1 + \lambda_2 + (k-1)(\lambda_1 + \beta_1)) P(S_2(t)) \quad (9)$$

$$\frac{\partial P(S_F(t))}{\partial t} = (k-1)\lambda_1 P(S_1(t)) + (k-1)(\lambda_1 + \beta_1) P(S_2(t)) \quad (10)$$

The reliability shows that only from state 0-2, the cluster will be in operational mode, and the data is also available. To derive the formula for mean time to data loss (MTTDL) for the cluster, considering the cluster will be staying at state0 – state2 and taking into that initial state of the cluster is state0 for calculating the MTTDL as follows:

$$MTTDL = \int_0^{\infty} P(S_0(t)) + P(S_1(t)) + P(S_2(t)) dt \quad (11)$$

By substituting the above values 7,8,9 and 10 in 11 we get,

$$MTTDL = \frac{H(x)}{V(x)} \quad (12)$$

Where as

$$H(x) = (\mu_D + k\lambda_0 + (k-1)\lambda_1)(\alpha_1 + (k-1)\beta_1) + (k\lambda_0 + (k-1)\lambda_1)(\mu_D + \lambda_2 + (k-1)\lambda_1)$$

$$V(x) = (k\lambda_0(k-1)\lambda_1)(\alpha_1 + \lambda_2 + (\mu_D(k-1)\lambda_1)(k-1)(\lambda_1 + \beta_1))$$

If the faulty node replacement rate  $\mu_D \rightarrow \infty$  then the simplified formula for MTTDL is

$$MTTDL = \frac{\alpha_1 + k\lambda_0 + (k-1)(\lambda_1 + \beta_1)}{k\lambda_0(k-1)(\lambda_1 + \beta_1)} \quad (13)$$

If the faulty node replacement rate  $\mu_D \rightarrow 0$  then the simplified formula is

$$MTTDL = \frac{1}{k\lambda_0} + \frac{1}{(k-1)\lambda_1} \quad (15)$$

Meantime to data loss occurs on the unavailability of data and failure of node capacity. MTTDL is decreasing when the node failure rate in the network has been increased.

#### IV. PERFORMANCE EVALUATION FOR DRAFT ALGORITHM

This section presents the simulation setup, performance metrics, comparison of performance analysis with and without the DRAFT algorithm incorporated into the network.

##### A. Simulation Setup

The proposed DRAFT algorithm has been executed in CUPCARBON. The simulation setup consists of 150 sensor nodes with 27 clusters and is randomly deployed in a 200 X 200m square region. Assuming that all sensor nodes are to be homogeneous resources and characteristics. Each node data size is 100 to 2000 units and generates the data items periodically. Every node in the cluster maintains a DNN attribute table which holds MAC address, neighbors list, node id and the storage space of each node learned through continuous resource management messages broadcast by every 10s. Each round simulation time is set to be 600s. Fault-tolerant system parameters are listed in the following Table III.

TABLE III. FAULT-TOLERANT SYSTEM PARAMETERS

| Fault-tolerant system parameters |                   |
|----------------------------------|-------------------|
| No. of Nodes                     | 100-150 Nodes     |
| Size of Data                     | 100 to 2000 units |
| Sensing interval                 | [1-5]s            |
| Broadcast interval               | 10s               |
| Round simulation time            | 600s              |
| Sensing field area               | 200 X 200m square |
| Node failure model               | Random            |
| MAC Protocol                     | 802.15.4          |
| Propagation loss model           | log distance      |
| Network Topology                 | Random            |

B. Performance Metrics

The impact of the DRAFT algorithm on an IoT network is analyzed through the following performance metrics.

1) *Network Lifetime*: In the simulation, the average Network lifetime of the IoT network with and without the DRAFT algorithm has been evaluated. This algorithm recovers the data when node failure occurs in the network. Considering ‘p’ is the probability of node failure that a node fails in one round. Assuming that the probability of a node failure for each round should vary from 0.1% to 0.5% as an increment. The total no. of deployed nodes in a cluster is ‘Ni’ and the communication range is ‘X’. For without recovery scheme, in around at least a single node failure probability is  $P_f$  and is  $1 - (1 - p)^{Nd}$  (Bernoulli’s trails) whereas Nd is the network density and considering as a standard value. In the case of with recovery scheme, ‘R’ is the recovery candidate, and there are two requirements to recover the data, first, the recovery candidate must be alive and the second is all of its direct neighbors should be alive. The probability of R is  $P_v = (1 - p)^{Nd}$  whereas  $P_v$  is the probability of a recovery candidate. The

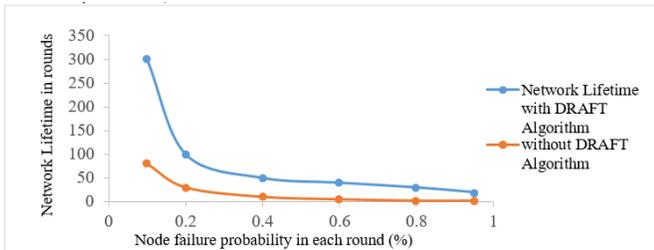


Fig. 4. Analysis of Network Lifetime with Respect to Node Failure Probability.

network lifetime of the proposed recovery scheme is more as compared to the without DRAFT algorithm as clearly shown in Fig. 4. When the probability of node failure is increased automatically the network lifetime is dropped for both with and without recovery scheme.

2) *Throughput*: Throughput is defined as the amount of data transmitted successfully from one node to another node in the network within a period. When the probability of the node failure decreases then the throughput of the network increases gradually as shown in Fig. 5. If there is a node failure occurs the data has been transmitted successfully because of the DRAFT algorithm. The throughput has improved double times with the recovery scheme as compared to the without recovery scheme.

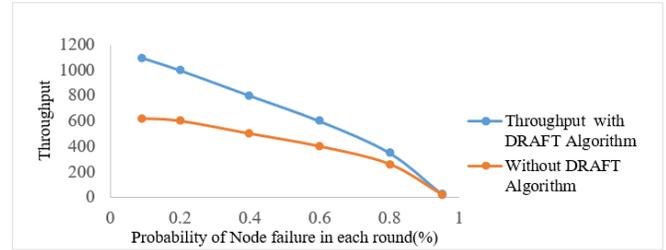


Fig. 5. Throughput with Respect to Probability Node Failure.

3) *MTTDL*: Meantime to data loss is the average time that causes data loss in the node. Data loss is occurred due to error situations in the networks. Backup and data recovery methods are helping to recover data or to avoid data loss in the IoT networks. If failure of any node in the cluster,

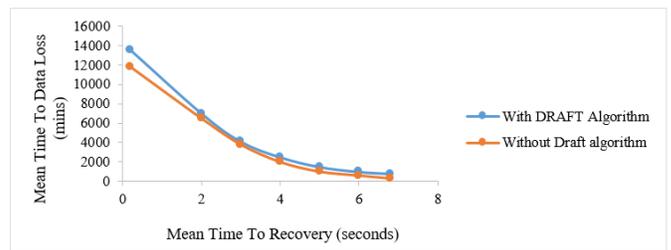


Fig. 6. Simulation Result of Mean Time to Data Loss.

MTTDL is decreased by increasing the recovery rate of the data. This will help to improve the network reliability and data availability in the network. The simulation results in Fig. 7 show the best recovery rate when the DRAFT algorithm has been incorporated into the network.

4) *Reliability*: Reliability is the capacity of the network to work during the presence of node failures concerning time. Here the time considering as the normalized time which means scaling the time within the range of 0 to 1. For an IoT network

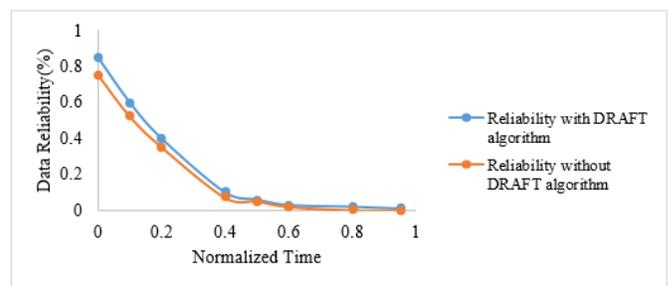


Fig. 7. Observations for Data Reliability in IoT Network.

at t = 0, the reliability is approximately high, with respect to time the network reliability is gradually decreasing which is shown in Fig. 6. The reliability mainly depends on the failure rate and repair rate of the node, and data recovery of the node.

## V. CONCLUSION AND FUTURE WORK

The proposed DRAFT algorithm is most suitable for IoT-based safety and critical applications. It is implemented with the concept of the RAID5 storage mechanism used for soft computing. Hence, this scheme can be easily incorporated with other IoT algorithms. By conducting a series of experiments the data reliability achieved as 85%. The throughput and network lifetime of the proposed algorithm are enhanced 5% as compared with the existing algorithms. The complete data recovery simulation is carried out which ensures reliable data transmission. In real scenario the data can be dropped due to noise, environmental factors, and data collisions. This paper presented analysis and simulation of simultaneous single IoT node failure in a cluster for the mentioned scenarios. This algorithm can be carry forward for the future multi-node data failures and data transmission reliability.

## REFERENCES

- [1] Hong Chen, David Hailey, Ning Wang, and Ping Yu. A review of data quality assessment methods for public health information systems. *International journal of environmental research and public health*, 11(5):5170–5207, 2014.
- [2] Xianke Sun, Gaoliang Wang, Liuyang Xu, and Honglei Yuan. Data replication techniques in the internet of things: a systematic literature review. *Library Hi Tech*, 2021.
- [3] Samaresh Bera, Sudip Misra, and Athanasios V Vasilakos. Software-defined networking for internet of things: A survey. *IEEE Internet of Things Journal*, 4(6):1994–2008, 2017.
- [4] Argyro Mavrogiorgou, Athanasios Kiourtis, Chrysostomos Symvoulidis, and Dimosthenis Kyriazis. Capturing the reliability of unknown devices in the iot world. In *2018 Fifth International Conference on Internet of Things: Systems, Management and Security*, pages 62–69. IEEE, 2018.
- [5] Mohamed Younis, Izzet F Senturk, Kemal Akkaya, Sookyoung Lee, and Fatih Senel. Topology management techniques for tolerating node failures in wireless sensor networks: A survey. *Computer Networks*, 58:254–283, 2014.
- [6] Chunsheng Zhu, Lei Shu, Takahiro Hara, Lei Wang, Shojiro Nishio, and Laurence T Yang. A survey on communication and data management issues in mobile sensor networks. *Wireless Communications and Mobile Computing*, 14(1):19–36, 2014.
- [7] Kun Wang, Yun Shao, Lei Xie, Jie Wu, and Song Guo. Adaptive and fault-tolerant data processing in healthcare iot based on fog computing. *IEEE transactions on network science and engineering*, 7(1):263–273, 2018.
- [8] Yuchang Mo, Liudong Xing, Wenzhong Guo, Shaobin Cai, Zhao Zhang, and Jianhui Jiang. Reliability analysis of iot networks with community structures. *IEEE Transactions on Network Science and Engineering*, 7(1):304–315, 2018.
- [9] Hui Yie Teh, Andreas W Kempa-Liehr, I Kevin, and Kai Wang. Sensor data quality: a systematic review. *Journal of Big Data*, 7(1):1–49, 2020.
- [10] Ericsson. Ericsson 50 billion connections. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 434–462. IEEE, 2020.
- [11] James Hong, Johnathan Raymond, and Joel Shackelford. Edisense: A replicated datastore for iot data. *Stanford University, Stanford*, 7(2):1–49, 2014.
- [12] Pietro Gonizzi, Gianluigi Ferrari, Vincent Gay, and Jérémie Leguay. Data dissemination scheme for distributed storage for iot observation systems at large scale. *Information Fusion*, 22:16–25, 2015.
- [13] Akshay Kumar, Nanjangud C Narendra, and Umesh Bellur. Uploading and replicating internet of things (iot) data on distributed cloud storage. In *2016 IEEE 9th International Conference on Cloud Computing (CLOUD)*, pages 670–677. IEEE, 2016.
- [14] Juan A Colmenares, Reza Dorrigiv, and Daniel G Waddington. A single-node datastore for high-velocity multidimensional sensor data. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 445–452. IEEE, 2017.
- [15] Waleed Bin Qaim and Öznur Özkasap. State-of-the-art data replication techniques in iot-based sensor systems. In *2018 IEEE Globecom Workshops (GC Wkshps)*, pages 1–6. IEEE, 2018.
- [16] Waleed Bin Qaim and Oznur Ozkasap. Draw: Data replication for enhanced data availability in iot-based sensor systems. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, pages 770–775. IEEE, 2018.
- [17] Chao Wang, Christopher Gill, and Chenyang Lu. Adaptive data replication in real-time reliable edge computing for internet of things. In *2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI)*, pages 128–134. IEEE, 2020.
- [18] Tzung-Shi Chen, Neng-Chung Wang, and Jia-Shiun Wu. An efficient adjustable grid-based data replication scheme for wireless sensor networks. *Ad Hoc Networks*, 36:203–213, 2016.
- [19] Berihun Fekade, Taras Maksymyuk, Maryan Kyryk, and Minh Jo. Probabilistic recovery of incomplete sensed data in iot. *IEEE Internet of Things Journal*, 5(4):2282–2292, 2017.
- [20] Yushi Shi, Xiaoqi Zhang, Qiaohong Hu, and Hongju Cheng. Data recovery algorithm based on generative adversarial networks in crowd sensing internet of things. *Personal and Ubiquitous Computing*, pages 1–14, 2020.
- [21] Chinmaya Mahapatra, Zhengguo Sheng, Victor CM Leung, and Thanos Stouraitis. A reliable and energy efficient iot data transmission scheme for smart cities based on redundant residue based error correction coding. In *2015 12th Annual IEEE International Conference on Sensing, Communication, and Networking-Workshops (SECON Workshops)*, pages 1–6. IEEE, 2015.
- [22] Teng Xu and Miodrag Potkonjak. Energy-efficient fault tolerance approach for internet of things applications. In *Proceedings of the 35th International Conference on Computer-Aided Design*, pages 1–8, 2016.
- [23] Chen Wang, Hoang Tam Vo, and Peng Ni. An iot application for fault diagnosis and prediction. In *2015 IEEE International Conference on Data Science and Data Intensive Systems*, pages 726–731. IEEE, 2015.
- [24] Sen Zhou, Kwei-Jay Lin, and Chi-Sheng Shih. Device clustering for fault monitoring in internet of things systems. In *2015 IEEE 2nd World Forum on Internet of Things (WF-IoT)*, pages 228–233. IEEE, 2015.
- [25] Enver Derun Karabeyoğlu and Tufan Coşkun Karalar. Iot module improves smart environment reliability. In *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*, pages 1–5. IEEE, 2017.

# An Enhanced Traffic Split Routing Heuristic for Layer 2 and Layer 1 Services

Ahlem Harchay, Abdelwahed Berguiga, Ayman Massaoudi  
Department of Computer Science  
Jouf University, Sakakah  
Saudi Arabia

**Abstract**—Virtual Private Networks (VPNs) have now taken an important place in computer and communication networks. A virtual private network is the extension of a private network that encompasses links through shared or public networks, such as the Internet. A VPN is a transmission network service for businesses with two or more remote locations. It offers a range of access speeds and options depending on the needs of each site. This service supports voice, data and video and is fully managed by the service provider, including routing equipment installed at the customer's premises. According to its characteristics, VPN has widely deployed on "COVID-19" offering extensive services to connect roaming employees to their corporate networks and have access to all the company information and applications. Hence, VPN focuses on two important issues such as security and Quality-of-Service. This latter has a direct relationship with network performance such as delay, bandwidth, throughput, and jitter. Traditionally, Internet Service Providers (ISPs) accommodate static point-to-point resource demand, named, Layer 1 VPN (L1VPN). The primary disadvantage of L1VPN is that the data plane connectivity does not guarantee control plane connectivity. Layer 2 VPN is designed to provide end-to-end layer 2 connection by transporting layer 2 frames between distributed sites. An L2VPN is suitable for supporting heterogeneous higher-level protocols. In this paper we propose an enhanced routing protocol based on Traffic Split Routing (TSR) and Shortest Path Routing (SPR) algorithms. Simulation results show that our proposed scheme outperforms the Shortest Path Routing (SPR) in term of network resources. Indeed, 72% of network links are used by the Enhanced Traffic Split Routing compared to Shortest Path Routing (SPR) which only used 44% of the network links.

**Keywords**—Virtual private network; enhanced traffic split routing; quality of service; shortest path routing; layer 1 VPN; layer 2 VPN

## I. INTRODUCTION

Computing, and networks in particular, have changed a lot over the past twenty years. The flow of information and the emergence of new technologies have increased considerably [1]. It is now possible to exchange substantial data of all types as well as to transmit voice and video over computer networks [2].

Indeed, with a modern economy based on new information and communication technologies, most companies use a set of means for the implementation of a reliable and flexible computer network [3]. This network allows corporate users to share resources such as printers, files and data. As a result, the need for remote connection to corporate resources has become common. Remote applications thus become the main tool of

the company's information system. The question that may arise then is how to ensure access within a structure sometimes spread over large geographical distances? In concrete terms, how can a branch of a company access data located on a server in the headquarters several thousand kilometers away?

Virtual private networks (VPNs) have been set up to respond to this type of problem and takes an important place in computer and communication networks. As pointed out by [4], a VPN is the extension of a private network that encompasses links through shared or public networks, such as the Internet.

It offers a range of access speeds and options depending on the needs of each site. This service supports voice, data and video and is fully managed by the service provider, including routing equipment installed at the customer's premises [5]. Indeed, the damaged caused by "COVID-19" on global economy leads company networks on looking for VPN solutions to establish a private communication to the corporate intranet while traveling from home.

Both service providers and customers are starting to realize the benefits of VPN solutions. New applications such as voice, telemedicine and video on demand make it possible to envisage an increase in productivity and a reduction in costs. However, VPNs are not only interested in extending LANs at a lower cost, but also in the use of specific services or functions ensuring quality of service (QoS) and security of exchanges [6] [7]. Indeed, the notion of quality of service makes it possible to formalize the requirements for each type of service in terms of performance criteria: bandwidth, end-to-end transmission delay, packet loss rate, jitter, etc. Each service may have different quality requirements.

Services such as voice or video impose very strong constraints on the quality of transmission: transmission delays or data loss must not degrade communication or the broadcasting of a video stream. In order to support real-time and multimedia applications on virtual private networks, it is necessary to develop routing algorithms which take QoS parameters into account. Routing algorithms with QoS must be adaptive and flexible for efficient management of resources in the network [8]. In practice, routing with QoS has not worked well.

The objective of routing is to determine a route (i.e. a set of links to be traversed), respecting certain constraints, to establish a connection from a source node to a destination node. The purpose of a routing algorithm is to allow the calculation of the route between those two nodes within the meaning of a certain criterion?

The remainder of the paper is organized as follows. The next section will focused on different algorithms such as Waxman and Brite algorithms. Section 3 highlights the approach proposed in this work. Then, an analysis of the performance of our prototype system implemented using simulation model will be described in Section 4. In Section 5, numerical results are presented to show the effectiveness of the proposed algorithm. Lastly, the conclusion and future work are outlined in Section 6.

## II. LITERATURE REVIEW

VPNs allow remote users, partners and providers to access certain parts of their networks (intranets). They also allow the deployment of many types of applications such as real-time voice or video, critical business management software or interactive applications [9] [10]. Originally, companies using VPN solutions used "layer 1" services such as "leased lines" (and referred to as layer 1 VPN, L1VPN). Leased lines are dedicated connections that a telecom operator operates directly between two customer sites, providing a permanent connection at a determined speed. Although leased lines offer users the confidentiality and reliability of transferred data, they suffer from a lack of flexibility compared to other types of layer 2 solutions such as Frame Relay, ATM, L2TP, L2F or also more recently Carrier/Metro Ethernet.

However, as mentioned in [11] [12], this type of VPN is characterized by its prolificity in singular domain and the lack the Quality of Service (QoS) during the inter-domain routing which lead to inhibit its scalability and flexibility. Another issue with L1VPN is the inter-configuration of a customer on another Service provider network as the policies are distinctives. This can be solved using the address mapping mechanism, unfortunately , this latter is not well-defined in standard specifications [13].

In fact, "layer 2" VPN services (or also layer 2 VPN, L2VPN) have allowed service providers to offer their customers a connection similar to that offered by leased lines. On the other hand, with L2VPN it is no longer necessary to have a dedicated leased line for each network interconnection. The clients share a single physical line and each has its own logical channels to send its traffic [14]. Layer 2 VPN services are attractive to the service operator because they do not require the operator to participate in the design and configuration of layers 2/3 of the customers' LAN. Also, the management and maintenance of the control plan are carried out by the customer and they are transparent to the operator's network [15] [16].

The major problem with L2VPN is security. Unlike L1VPN where each customer has their own private line, a layer 2 VPN is deployed on a shared network infrastructure that can be managed by national and international network service providers. As a result, several companies disagree that their data should be transferred through shared, unsecured tunnels [17]. One solution would be to offer a layer 3 or L3VPN VPN service. Security is usually provided by a combination of tunneling and encryption methods. The best known is the one that implements the IPSec (IP Security) protocol. IPSec is a "layer 3" security protocol. It is based on the IP protocol and offers tunneling and security features, including encryption, authentication and key management [18]. In this

work, a random generator graph named "Brite" is used [19]. The BRITE topology generator assigns each link with a delay based on its physical distance. The algorithm can be described as follows [20]:

- Firstly, we specify the number of nodes on the networks.
- For a link creation between two nodes  $u$  and  $v$  we define the probability  $P(u, v)$  :

$$P(u, v) = \beta \exp \frac{-d(u, v)}{L\alpha} \quad (1)$$

Where,

- $d(u, v)$ : the distance separating from node  $u$  to node  $v$ ;
- $L$ : the maximum distance between node  $u$  and node  $v$ ;
- $\alpha$  and  $\beta$ : These two constant parameters are defined in the interval (0.1).

When the constant  $\alpha$  is decreased, we noticed that the link's density on the network is increased. Based on the link probability  $(u, v)$  a link is added or not between  $u$  and  $v$ . In a shortest path tree problem, we consider a directed graph  $G = (V, E)$ , where  $V$  represents the set of nodes and  $E$  the set of links. Each edge has a weight  $P_i$ . A path  $C = \langle e_1, e_2, \dots, e_n \rangle$  has a weight which represents the sum of the weights of the edges constituting the path. The shortest path from a vertex  $d$  to a vertex  $a$  is the minimum weight path that connects  $d$  to  $a$  [21].

The two algorithms Bellman-Ford and Dijkstra described in [22] are two well known shortest path algorithms. Shacham [23] proposed a maximum bandwidth tree algorithm to distribute data hierarchically. It uses an algorithm close to Dijkstra to calculate the maximum bandwidth of a single path to all destinations.

The principle of this algorithm is as follows:

- 1) Determine the maximum bandwidth paths available between the different nodes.
- 2) Sort the receivers according to their reception capacities.
- 3) Add recipients to the maximum bandwidth tree one by one.

This hierarchical distribution approach gives for each individual receiver the rate at which it will receive data from the source. The bandwidth will then be allocated appropriately.

## III. PROPOSED ALGORITHM

As aforementioned, there are many issues on deploying Dijkstra and Bellman-Ford for routing protocols using two or more constraints. Dijkstra as well Bellman-ford are deployed where we using one constraint and offer good paths. However, when we need the formation of a balanced system when distributing the load this is not guarantee by Dijkstra and Bellman-ford which use an order of priority in the constraint's choice.

Our proposed scheme, named Enhanced Traffic Split Routing (E-TSR), is based on algorithm Traffic Split Routing (TSR) [24] [25] which offer a good objective on load balancing inside a network. Enhanced Traffic Split Routing try to distribute homogeneously the traffic on the network and offer a balanced sharing of traffic. Indeed, with E-TSR, the maximum possible of links are used to balance the traffic on the network.

We present in what follows, *algorithm 1*, the heuristic of traffic distribution used which is an enhanced algorithm based on [24] [25]. To begin, it is interesting to note that is not always optimal to use the shortest path between a pair of nodes "i" and "j". Accordingly, we will use a model of an M/M/1 queue. Suppose that between two nodes "i" and "j" we have two paths: the shortest path of length "n" and another longer path of length  $m > n$ .

To begin with, we assume that we have a first path calculated by the shortest path algorithm. This path links a source "i" to a destination "j" and uses "n" links. We assume that each link in the path from "i" to "j" is modeled by an independent M/M/1 queue (*Kleinrock independence assumption*) as illustrated on Fig. 1.

Now suppose that the traffic is shared between the path with "n" hops and that of "m" hops. Consider the following variables:

$\rho$ : use of the link,  
when all the traffic is offered to the first path only (the shortest path) we have the average residence time on a link:

$$T = \frac{1}{\mu - \lambda} \quad (2)$$

Therefore, the use of the link

$$\rho = \frac{\lambda}{\mu} \quad (3)$$

Since this path is composed of "n" independent links, the average residence time in the path is modeled by the following formula:

$$T_1 = n \times T = \frac{n}{\mu - \lambda} \quad (4)$$

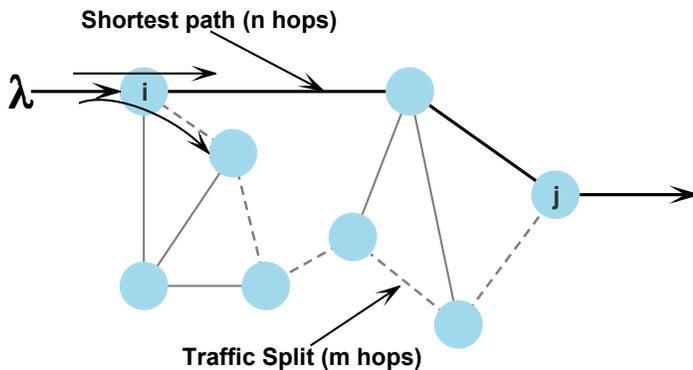


Fig. 1. Shorter Path Versus Sharing of Traffic Through Disjointed Paths.

If we share the traffic between the two paths; the first made up of "n" hops and the second of "m" hops, we have the variables:

- $\lambda_1$ : the arrival rate for path 1

- $\lambda_2$ : the arrival rate for path 2

So, the average residence time across the two paths is:

$$T_2 = \frac{\lambda_1}{\lambda} \frac{n}{\mu - \lambda_1} + \frac{\lambda_2}{\lambda} \frac{n}{\mu - \lambda_2} \quad (5)$$

If we find cases where  $T_1 < T_2$  then we can state that the shortest path does not give precisely the best delay. We can say that:

$$T_1 \leq T_2 \Leftrightarrow \frac{\rho_1}{1 - \rho_1} + \frac{m}{n} \frac{\rho_2}{1 - \rho_2} \leq \frac{\rho}{1 - \rho} \quad (6)$$

Where,

$$\rho_1 = \frac{\lambda_1}{\mu} \quad (7)$$

$$\rho_2 = \frac{\lambda_2}{\mu} \quad (8)$$

$$\rho \geq \rho_1 \quad (9)$$

$$\rho \geq \rho_2 \quad (10)$$

$$\rho = \rho_1 + \rho_2 \quad (11)$$

In the case where  $m < n$ , then the inequality 6 gives us:

$$m \leq \frac{n(1 - \rho_2)}{(1 - \rho_1)(1 - \rho)} \quad (12)$$

The inequality 12 shows that the number of hops in the path should be small and not greater than a certain constant. Additionally, this inequality clarifies that when the shorter path is overloaded (maximum link utilization), using another longer path to route traffic can be useful in order to reduce the wait time. Moreover, we can deduce that the number of hops in the longest path decreases when the load offered to this path ( $\rho_2$ ) increases. In particular, in the case where the traffic is distributed between the shortest path and the longest one ( $\rho_1 = \rho_2$ ), it suffices to have  $m < n/((1 - \rho))$  to reduce the waiting time by traffic distribution. For example, if  $\rho = 80\%$ , then we must have  $m < 5n$ . That is, the number of hops in the longest path should not exceed 5 times the number of hops in the shortest path.

#### IV. RESULT AND DISCUSSION

This section is devoted to evaluating the performance and quality of services resulting from the Enhanced Traffic Split Routing (E-TSR) algorithm by comparing the results with those obtained with Traffic Split Routing (TSR) [25] and the Shortest Path Routing (SPR). In order to be able to evaluate these three algorithms, various simulations were carried out using the NS-2 simulation platform [26].

To begin, we attempt to give an overview of different parameters related to evaluate the performance of our proposed scheme. Then, we will present more closely the NS-2 tool as

**Algorithm 1** Enhanced Traffic distribution heuristic

**Input** :  $L_s$  the number of times a link  $S$  appears in a VPN tree

- 1: **procedure** HEURISTIC PROCEDURE
- 2:      $L_s \leftarrow 0$  :Definition of a link Variable
- 3:      $n \leftarrow 0$  :Definition of the number of nodes on the network
- 4:     *loop*: waiting for a new Virtual Private Network connection demand from any network's node
- 5:     Complete (or Generate) a path (tree) coupling all the new Virtual Private Network and avoiding links whose  $L_s > n$
- 6:     **if** path is icomplete **then**
- 7:          $n \leftarrow n + 1$  and **goto** step 5
- 8:     **else**
- 9:          $L_s \leftarrow L_s \log(L_s) + 1$  for all the links of the new generated tree and **goto** step 3.
- 10:    **end if**
- 11: **end procedure**

well as the network model used to perform different scenarios to be simulated under NS-2.

The rest of this section will highlights the QoS parameters used to evaluate the three heuristics: the enhanced traffic distribution (TSR, Traffic Split Routing), the traffic distribution (TSR, Traffic Split Routing), and the shortest path (SPR, Shortest Path Routing). All the simulation parameters are given in Table I. Our simulations are performed using the

TABLE I. SIMULATION PARAMETERS

| Parameters                               | Values       |
|------------------------------------------|--------------|
| Simulator name                           | NS-2         |
| Node's number                            | 24 nodes     |
| Tree's link capacity                     | 100 Mb/s     |
| Transmission delay                       | 10 ms        |
| Source's number of the generic tree      | 4-24 nodes   |
| Simulation Time                          | 2000 seconds |
| Application type used on the simulations | FTP          |
| Packet size                              | 1 KB         |

NS-2 network simulator. For accuracy and compliance, all simulations are performed Twenty times for each scenario. All simulations are performed to study the behavior of the three routing algorithms; E-TSR, TSR, and SPR. All simulations are generated with different random number seeds and the results are averaged over all the outcomes. Fig. 2, Fig. 3, and Fig. 4 illustrate an example of network scenario used in performance study for the three heuristics E-TSR, TSR, and SPR.

Fig. 4 illustrates the scenario using Shortest Path traffic algorithm. From this figure, we suppose that the traffic is sent from node 2 to the destination node 4. We see that the node named 6 used as a Steiner node and all traffic is focused on the shortest path (from node 2 to node 6, and from node 6 to node 4). On the other hand, Fig. 2 illustrates the scenario of the Enhanced Traffic Split Routing. From this Figure we can see that the traffic is shared equitably between different paths. Indeed, the traffic sent from node 2 and has as destination node 4 has taking different paths (from node 2 to node 6, from node 2 to node 7, etc.). with this approach, we can see that we use maximum links on the network.

*A. Average Reception Data Rate*

Fig. 5 illustrates the average data rate reception of the three algorithms, E-TSR, TSR and SPR. As we can see, the Fig. 5 shows the average data rate as a function of the number

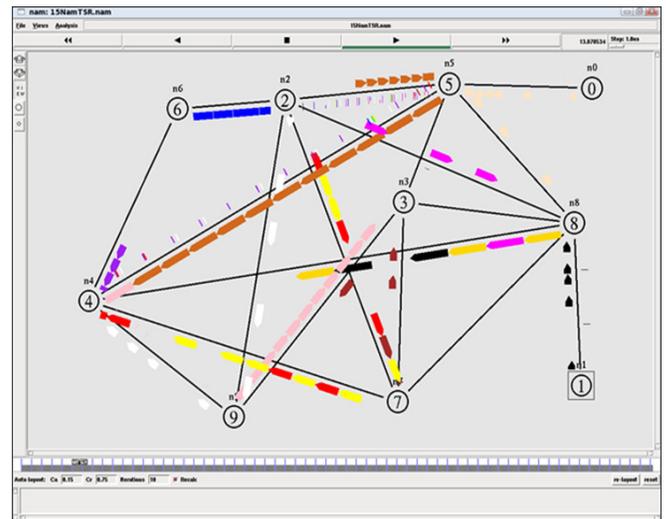


Fig. 2. Enhanced TSR Traffic.

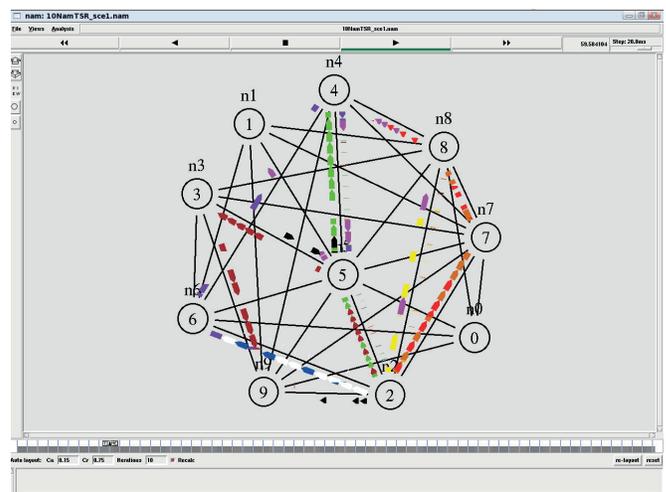


Fig. 3. TSR Traffic.

of source VPNs. Indeed, in the case of 6 source VPNs we have 5.9 Mbps with E-TSR heuristic, 5.58 Mbps with TSR heuristic while the average throughput with SPR is 5.23 Mbps.

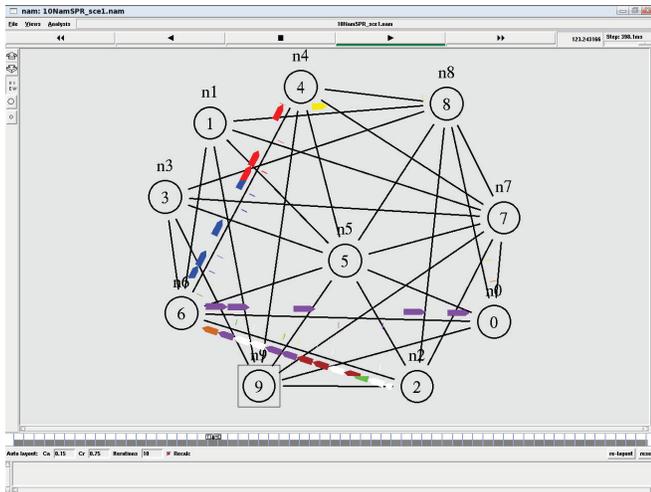


Fig. 4. SPR Traffic.

subsequently, with 16 source VPNs, the average throughput with E-TSR is equal to 5.35 Mbps while it is equal to 3.93 Mbps with SPR algorithm. We see also that the gap on the average data rate increase with the number of source VPNs and the E-TSR algorithm offer good throughput compared to TSR and SPR algorithms. This is due to algorithm properties. Indeed, with SPR, all the traffic is focused on the shortest path while with E-TSR the traffic is divided between the maximum number of links on the network. Moreover, the use of maximum links allowed networks to offer a higher throughput especially for networks with large traffic, unlike to shortest path algorithm which focused only on some links (shortest link) which lead to some links to become overloaded, leaving others unused. This had an influence on the flow and then on the error rate.

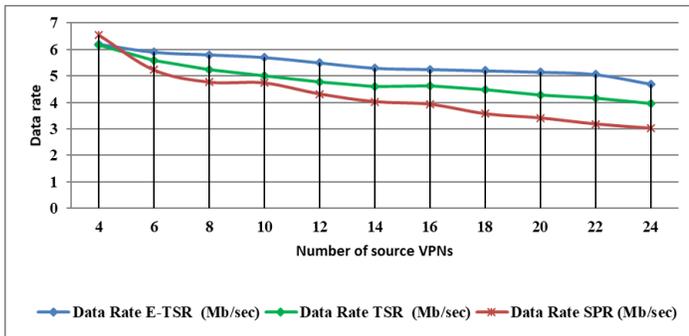


Fig. 5. Average Data Rate Reception.

### B. Packet Loss Rate

Fig. 6 illustrates the packet loss rate as a function of number of source VPNs for each routing technique E-TSR, TSR and SPR. As shown in this figure, with low number of source VPNs average error rate with E-TSR and TSR become more frequent due to the fact that the routing techniques must search new link every time there is a new source VPN. It can be noticed that when the number of VPN sources increases

to 12 sources, SPR and E-TSR have almost the average error rate,  $9.6 \times 10^{-4}$  and  $11.5 \times 10^{-4}$ , respectively.

This rate increases to reach  $43 \times 10^{-4}$  packets loss with SPR routing technique,  $30 \times 10^{-4}$  packets lost with TSR routing technique, and  $27 \times 10^{-4}$  packets lost with E-TSR routing heuristic for 24 source VPNs.

Furthermore, as the number of source VPNs increase, the gap between SPR, TSR, and E-TSR increases and as we can see E-TSR provides less packet error rate. Indeed, with a shortest path routing technique all the traffic takes the same path which lead to a huge traffic on some links and then more error rate. However, with distributed traffic routing technique the traffic flow is sent over the network moderately over all links.

Furthermore, using a traffic distributed technique, packet have low chance to enter on overloaded queues, thus dropping packets will minimized and rejecting packets will be decreased. On the other site, with shortest path routing technique, there is a high probability the traffic takes the same path leading to a queue overload and then increase rejected packets.

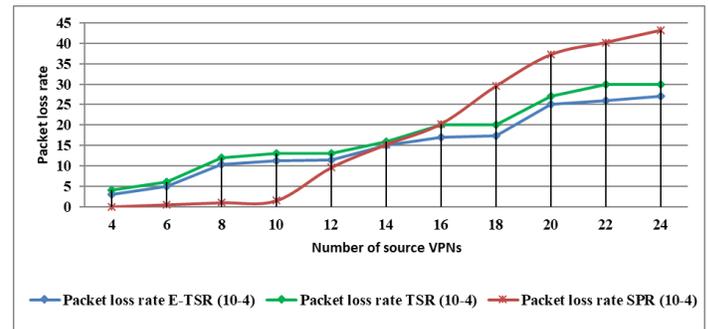


Fig. 6. Average Packet Loss Rate.

### C. Average End-to-end Delay

In this section we examine the mean delay to send a packet from one source to a destination. This delay is given according to the number of source VPNs on the network. Fig. 7 shows that the delay of three heuristics E-TSR, TSR, and SPR increases linearly with the number of source VPNs.

Moreover, Enhanced Traffic Split Routing algorithm exhibit a brief variation on delay compared to the Shortest Path Routing Algorithm. From Fig. 7 we remark that SPR has a delay around of 30 ms with 6 source VPNs. By increasing the number of source VPNs we see that E-TSR offer less end-to-end delay compared to TSR and SPR. In fact, E-TSR and TSR look increase on logarithmic fashion compared to SPR.

It is clear to see that in the case of 16 source VPNs the average end-to-end delay is equal to 25.32 ms for E-TSR algorithm whereas it is equal to 27 ms for TSR algorithm, and 39 ms for SPR algorithm. Table II gives an overview of the measured values related to the mean delay for three routing algorithms. The gap on the end-to-end delay increases as number of source VPN increase. Our observations, for instances imply that when the network deploy a traffic distributed technique offer more chance of using a less queue memory which means packets

TABLE II. GAP MEAN DELAY

| # VPNs | Mean delay SPR (ms) | Mean delay TSR (ms) | Mean delay E-TSR (ms) |
|--------|---------------------|---------------------|-----------------------|
| 4      | 24                  | 25                  | 23                    |
| 6      | 30                  | 26                  | 23.56                 |
| 8      | 34                  | 26                  | 23.58                 |
| 10     | 36                  | 26                  | 23.99                 |
| 12     | 37                  | 26                  | 24                    |
| 14     | 39                  | 26                  | 24.12                 |
| 16     | 39                  | 27                  | 25.32                 |
| 18     | 40                  | 27                  | 25.64                 |
| 20     | 40                  | 27                  | 25.87                 |
| 22     | 41                  | 27                  | 25.9                  |
| 24     | 41                  | 28                  | 26                    |

sent on different links have more chance of going through small queues and therefore a small delay variation as shown on Fig. 8.

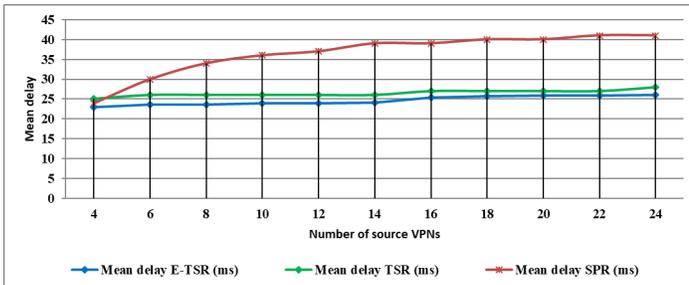


Fig. 7. Mean Delay.

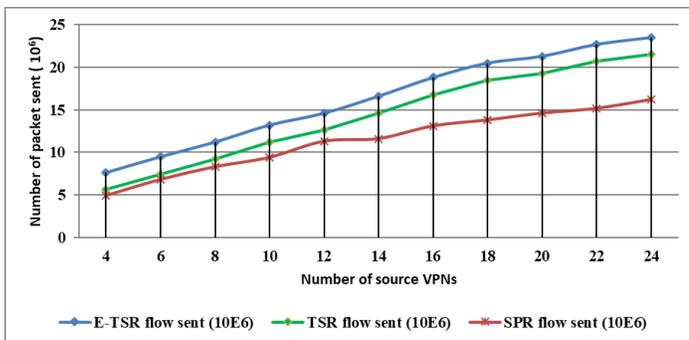


Fig. 8. Flow of Sent Data.

## V. CONCLUSION AND FUTURE WORK

In this paper, we proposed an enhanced traffic split routing algorithm. Such algorithm is compared with shortest path algorithm. Simulations were performed using NS-2 to analyze the functionality and performance of the proposed algorithm in terms of average data rate, packet loss rate, and average end-to-end delay.

The results show that Enhanced Traffic Splitting Routing algorithm provides least values on packet loss rate and average end-to-end delay compared to Shortest Path Routing and legacy Traffic Splitting Routing algorithm.

Also, simulation results show that TSR provides better performance in term of average data rate. So, it is concluded that enhanced traffic split routing algorithm has the capability

to provide better low packet loss rate and data rate by using 72% of network links compared to shortest path routing algorithm which uses only 44% of network links.

As a future work, we plan to design and implement the proposal experimentally in order to study these factors practically and exploring the potential of utilizing enhanced traffic split routing on real-time multimedia and VoIP applications.

## REFERENCES

- [1] R. M. Hicks, "Plan for always on vpn," in *Implementing Always On VPN*. Springer, 2022, pp. 7–20.
- [2] H. H. Elkarash, N. M. Elshennawy, and E. A. Saliem, "Evaluating qos using scheduling algorithms in mpls/vpn/swimax networks," in *2017 13th International Computer Engineering Conference (ICENCO)*. IEEE, 2017, pp. 14–19.
- [3] A. Black, T. Bui, S. Jenni, V. Swaminathan, and J. Collomosse, "Vpn: Video provenance network for robust content attribution," in *European Conference on Visual Media Production*, 2021, pp. 1–10.
- [4] T. Alam and K. Hamid, "Implementation of dynamic multipoint vpn over ipsec for secure enterprise network," Ph.D. dissertation, IUC Central Library, 2018.
- [5] A. BAHNASSE and N. E. KAMOUN, "Policy-based automation of dynamique and multipoint virtual private network simulation on opnet modeler," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 12, 2014.
- [6] F. Bensalah and N. El Kamoun, "Novel software-defined network approach of flexible network adaptive for vpn mpls traffic engineering," *Int. J. Adv. Comput. Sci. Appl*, vol. 10, no. 4, pp. 280–284, 2019.
- [7] M. Iqbal and I. Riadi, "Analysis of security virtual private network (vpn) using openvpn," *International Journal of Cyber-Security and Digital Forensics*, vol. 8, no. 1, pp. 58–65, 2019.
- [8] T. Vitalii, B. Anna, H. Kateryna, and D. Hrebeniuk, "Method of building dynamic multi-hop vpn chains for ensuring security of terminal access systems," in *2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T)*. IEEE, 2020, pp. 613–618.
- [9] Q. Jin, Q. Guo, M. Luo, Y. Zhang, and W. Cai, "Research on high performance 4g wireless vpn for smart factory based on key technologies of 5g network architecture," in *2020 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 2020, pp. 1443–1447.
- [10] Y. Christov, "Building personal virtual private networks in public cloud platforms," *Industry 4.0*, vol. 5, no. 3, pp. 112–113, 2020.
- [11] A. Bahnsasse, M. Talea, A. Badri, F. E. Louhab, and S. Laafar, "Smart hybrid sdn approach for mpls vpn management on digital environment," *Telecommunication Systems*, vol. 73, no. 2, pp. 155–169, 2020.
- [12] K. Gaur, A. Kalla, J. Grover, M. Borhani, A. Gurtov, and M. Liyanage, "A survey of virtual private lan services (vpls): Past, present and future," *Computer Networks*, p. 108245, 2021.
- [13] T. Takeda, R. Aubin, M. Carugi, I. Inoue, and H. Ould-Brahim, "Framework and requirements for layer 1 virtual private networks," RFC 4847, April, Tech. Rep., 2007.
- [14] S. M. Rosu, M. M. Popescu, G. Dragoi, and I. R. Guica, "Virtual enterprise network based on ipsec vpn solutions and management," *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 11, 2012.
- [15] S. T. Aung and T. Thein, "Comparative analysis of site-to-site layer 2 virtual private networks," in *2020 IEEE Conference on Computer Applications (ICCA)*. IEEE, 2020, pp. 1–5.
- [16] B. Wen, G. Fioccola, C. Xie, and L. Jalil, "A yang data model for layer 2 virtual private network (l2vpn) service delivery," *Internet Eng. Task Force, Fremont, CA, USA, Rep. RFC*, vol. 8466, 2018.
- [17] K. Arai, "Routing protocol based on floyd-warshall algorithm allowing maximization of throughput," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 6, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0110655>

- [18] H. Gunleifsen, T. Kemmerich, and V. Gkioulos, "Dynamic setup of ipsec vpns in service function chaining," *Computer Networks*, vol. 160, pp. 77–91, 2019.
- [19] B. Waxman, "Routing of multipoint connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, 1988.
- [20] E. Akin and T. Korkmaz, "An efficient binary-search based heuristic for extended unsplittable flow problem," in *2017 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2017, pp. 831–836.
- [21] F. U. Islam, G. Liu, and W. Liu, "Identifying voip traffic in vpn tunnel via flow spatio-temporal features," *Mathematical Biosciences and Engineering*, vol. 17, no. 5, pp. 4747–4772, 2020.
- [22] R. L. R. Thomas H. Cormen, Charles E. Leiserson and C. Stein, "Introduction to algorithms," 1997.
- [23] N. Shacham, "Multipoint communication by hierarchically encoded data," in *[Proceedings] IEEE INFOCOM '92: The Conference on Computer Communications*, 1992, pp. 2107–2114 vol.3.
- [24] A. Meddeb, "Benefits of multicast traffic split routing in packet switched networks," in *2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577)*, vol. 4, 2004, pp. 2019–2023 Vol.4.
- [25] A. Berguiga, A. Harchay, A. Massaoudi, and R. Khdir, "A new traffic distribution routing algorithm for low level vpns," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020.
- [26] K. R. Fall and K. Varadhan, "The ns manual (formerly ns notes and documentation)," 2002.

# AuSDiDe: Towards a New Authentication System for Distributed and Decentralized Structure based on Shamir's Secret Sharing

Omar SEFRAOUI<sup>1</sup>, Afaf Bouzidi<sup>2</sup>, Kamal Ghoumid<sup>3</sup>  
National School of Applied Sciences  
Engineering Sciences Laboratory LSI  
Oujda, Morocco

El Miloud Ar-Reyouchi<sup>4</sup>  
Department of Telecommunication and Computer  
Science, Abdelmalek Essaadi University  
Tétouan, Morocco

**Abstract**—Nowadays, connected devices are growing exponentially; their produced data traffic has increased unprecedentedly. Information systems security and cybersecurity are critical because data typically contain sensitive personal information, requiring high data protection. An authentication system manages and controls access to this data allowing the system to ensure the legitimacy of the access request. Most of the current identification and authentication systems are based on a centralized architecture. However, some concepts as Cloud computing and Blockchain use respectively distributed and decentralized architectures. Users without a central server will own platforms and applications of the next generation of Internet and Web3. This paper proposes AuSDiDe, a new authentication system for the distributed and decentralized structure. This solution aims to divide and share keys toward different and distributed nodes. The main objective of AuSDiDe is to securely store and manage passwords, private keys, and authentication based on the Shamir secret sharing algorithm. This new proposal significantly reinforces data protection in information security.

**Keywords**—Shamir's secret sharing; authentication system; decentralized; distributed; blockchain

## I. INTRODUCTION

Currently, the number of connected machines is growing exponentially. This is explained by several factors such as the use of social networks, streaming and sharing videos, online services (payment, purchases, etc.), connected objects, cryptocurrency [1].

The world has never been digitized as it is today. Digitization of different services, the use of cryptocurrency, IoT, etc. Moreover, the Covid-19 pandemic has given a boost to digitization, paving the way for new opportunities for growth, competitiveness and inclusion. The global data traffic has increased at an unprecedented rate over the last decade. There are various challenges and issues associated with this data.

Among these challenges, security and privacy have been considered as important issues since data often involves different types of sensitive personal information, e.g., addresses, personal preference, banking details, governmental data, financial, medical, military, or NFT's - Non-fungible token. Putting the data on the Net and on the Cloud [2] makes them vulnerable. This requires more vigilance from the administrators and owner of this data. To put more reliable and secure means to protect and secure the data against malicious people

and botnets. It is necessary to choose the security criteria to be taken into account. Commonly used security criteria are availability, integrity, and confidentiality, but it may be relevant to add others such as proof, control, anonymity, reliability. The scale of needs will be determined according to these security criteria. Information systems security and cybersecurity ensure data protection. An authentication system manages and controls access to this data [3]. Fig. 1, 2, and 3 show a centralized, decentralized, and distributed organizational structure.

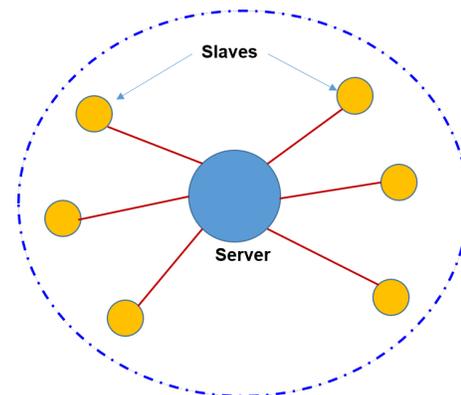


Fig. 1. Centralized Organizational Structure.

In Fig. 1, server systems with one or more slave nodes directly linked to a central server are centralized systems. In many companies, this is the most frequent sort of system.

In Fig. 2, every node in a decentralized system makes its own choice. The sum of the individual node choices determines the system's ultimate behavior. It is worth noting that the request is not received and responded to by a single organization structure.

A distributed system shown in Fig. 3 comprises a group of loosely coupled processors that are linked together via a communication network. A distributed system may consist of computational and diverse nodes connected by a communication network. Any node's total resources should be visible and freely available to other nodes. The choice of a load sharing or global planning technique is an important aspect of a distributed system's design configuration.

The advantage of the proposed approach, on the contrary

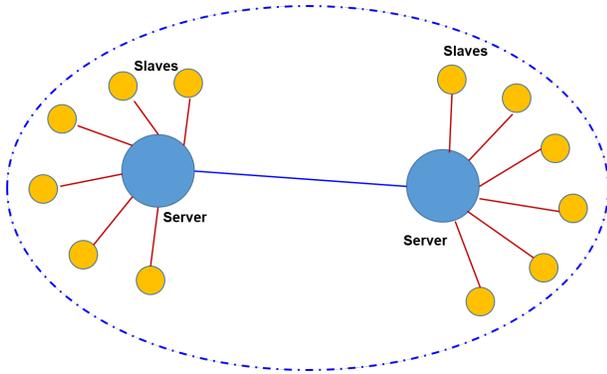


Fig. 2. Decentralized Organizational Structure.

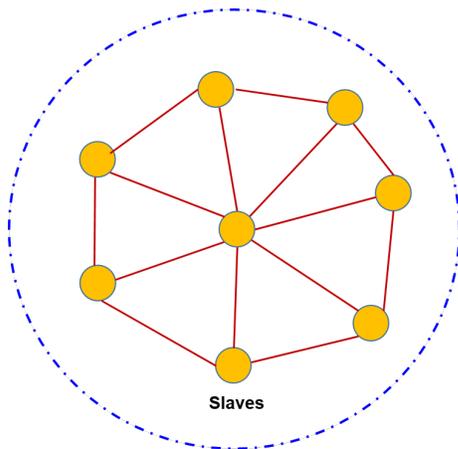


Fig. 3. Distributed Organizational Structure.

to most traditional secret sharing schemes, is that the shares are distributed and decentralized. This study presents a novel authentication scheme for the distributed (Fig. 3) and decentralized (Fig. 2) structure. This method aims to split and share keys across several remote nodes.

Most of the current identification and authentication systems are based on centralized architecture [4], today it is necessary to think about new methods and structures in order to strengthen the authentication and to be compatible with the new architectures.

There are some concepts like Cloud computing [5] and Blockchain [6] use respectively distributed and decentralized architectures [7] as shown in Fig. 2 and 3. For example, in the next generation Internet, the platforms and applications built on the Web3 will be owned by users, without a central server [8]. In order to remedy these limitations and security concerns, this paper presents a new approach, focused on the authentication systems, key and password management.

In Docker, integrity comes from trust, i.e. the user decides to retrieve a Docker image from the Docker hub, which is a set of community images. Docker security relates to threats and attacks that challenge security based on confidentiality, integrity, and availability of services. The application system must function flawlessly during the expected use ranges and guarantee access to the services and resources installed with

the expected response time. Basic Docker provides functionality to have high availability, Docker Swarm, an orchestrator that allows you to manage your cluster of containers. Indeed, the security of the information system is considered one of the primary issues to be established. Each organization must define a security policy to succumb to its needs and protect these resources and their trade secrets.

In order to increase the level of security [9], a new authentication mechanism is introduced. This authentication mechanism for distributed and decentralized structures is based on Shamir's secret sharing.

Hadoop is a big data processing paradigm that can efficiently address the issues of big data because of its distributed storage and parallel processing properties, as well as other benefits such as open source. The proposal is considered as the last method to guarantee the robustness of any key management system is to divide the keys into various bits, as recommended in the Split Keys approach. In this approach, no one individual has access to the real key; instead, the key must be used by a group of people. The system suggested in this paper splits and distributes keys to distinct and scattered nodes in all clusters.

AuSDiDe a new proposed architecture to securely manage passwords, private key and authentication. Shamir's algorithm is used to share secret information

The main purpose of this system is to split the secret key into parts, giving each server its own shared key, where some or all of the parts are needed in order to rebuild a passphrase that gives access to the secret.

The remainder of the paper is laid out as follows. First a review of the related work realized in this topic is presented. Then a presentation of the architecture of an authentication system. After the implementation results are displayed. The conclusions and future work are presented in the last Section.

## II. RELATED WORK

Nowadays, there are many proposals for securing and managing authentication systems. The proposed system reinforces security, prevents certain attacks such as man in the middle, identity theft, and adapts to new distributed and decentralized architectures.

To improve the security of keys used for encryption, [10] proposes a threshold secret sharing system employing Newton division difference interpolating polynomial in a distributed Cloud context.

The study [11] proposes a system that employs secure multiparty computation (SMPC) protocols with Shamir secret sharing for password- and iris-based authentication.

Hashing may not be able to hide data as effective in post quantum era [12], an authentication protocol which will use Shamir's secret sharing method to authenticate with server is proposed. A novel approach based on blockchain technology [13], digital signatures and threshold ElGamal Cryptosystem to address the problem of single point of failure.

The authors in [14] provide a viable way for protecting the traditional password-based authentication system since this

sort of authentication is often required. They suggest a way for sharing a secret based on Shamir's well-known (k, n) threshold approach.

In [15], the authors discuss a new approach for managing the secrets in a decentralised way by leveraging decentralised identity concepts such as verifiable credential technologies, password-authenticated key exchange protocols and multi-party computation.

Another approach were developed in [16] who introduce PASSAT is a practical method that enhances the security assurance provided by today's cloud architecture without needing any modifications or collaboration from cloud service providers. PASSAT is a cloud-invisible program that enables users to safely and effectively save and retrieve their files on public cloud storage with a single master password.

### III. SHAMIR'S SECRET SHARING

Authentication for a computer system is a process that allows the system to verify the legitimacy of the access request. The system then assigns this entity the identity data for this session. Access to the resources of an information system by an entity is broken down into three sub-processes, authentication, identification and access control [17].

Cryptography is one of the disciplines of cryptology focusing on protecting messages ensuring confidentiality, authenticity and integrity by often using secrets or keys. Shamir's Secret Sharing is an example of this cryptographic algorithm [18].

#### A. Distributed Scenario

A distributed system is a computer environment in which numerous nodes on a network are used to distribute different components [7].

Because distributed networks are formed up of equal, interconnected nodes, data ownership, and computing resources are dispersed equitably throughout the network. These nodes divided up the work and coordinated their efforts to finish the assignment more quickly than if it had been assigned to a single node. Compared to decentralized networks, distributed networks are more scalable.

A node in a distributed network may fail on its own without impacting the rest of the system. It is more difficult to change information on a distributed network since data is spread equitably throughout the whole network [7].

#### B. Decentralized

Decentralization is defined by the distribution of powers. Scaling decentralized networks is simple. Instead of depending on a single central node, a decentralized network spreads information processing among numerous nodes [7]. Even if one of the master nodes fails, the remaining servers can continue to offer data access to users, and the network as a whole will keep running. Because information kept on the network is spread across numerous points rather than via a single one, decentralized networks provide a higher level of consumer privacy [19]. Furthermore, user requests are often completed faster when using a decentralized network. Fig. 4 shows the nodes of the AuSDiDe node's topology.

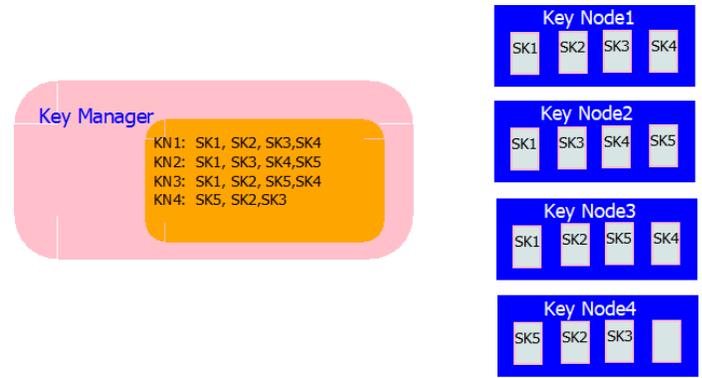


Fig. 4. AuSDiDe Nodes.

Fig. 5 details a schematic diagram of the general AuSDiDe architecture used for Distributed and Decentralized Structure Based on Shamir's Secret Sharing.

#### C. Shamir's Secret-Sharing Scheme

Shamir's Insider Information Adi Shamir is the creator of sharing. A secret is split into pieces in a kind of dispersed secret [18]. Each participant has their own shared key, where some or all the parts are needed to reconstruct a passphrase. Shamir's secret sharing or key sharing, is a process which a private encryption key is split into separate fragments. Every fragment is useless, unless it is sufficiently assembled to reconstitute the original key [7].

It is not necessary that all the participants reconstitute the access password. This is why the threshold scheme is sometimes used where a number  $k$  of the parts is sufficient for rebuild the original secret [20].

In Shamir's secret-sharing arrangement, there are  $n$  shareholders [21].  $U_i = \{U_1, U_2, \dots, U_n\}$  and a mutually trusted dealer  $D$ .

The dealer  $D$  produces a  $(t-1)$  degree polynomial  $f(x) \in Z_p$ , where  $P$  is a prime integer, to divide the secret  $S$  into  $n$  shares.  $S = f(0)$  is the shared secret, and before sending the  $(x_i, y_i)$  to the shareholder  $U_i$  the dealer calculates the secret-sharing shares as  $y_i = f(x_i)$  for  $x_i \neq 0$ .

When regenerating the secret, at least  $t$  shares  $(x_i, y_i)$  are required to recover the polynomial  $f'(x)$ , allowing each shareholder to get the secret  $S = f'(0)$ . The approach is comprised of two algorithms: secret reconstruction and share generation [21]:

For share generation, the  $(t-1)$  degree polynomial is defined as:

$$f(x_i) = a_0 + a_1x^1 + ax^2 + \dots + a_{t-1}x^{t-1}, (mod p) \quad (1)$$

and  $a_i \in Z_p$ , for  $0 \leq i \leq t-1, a_{t-1} \neq 0$ , the secret  $S = f(0) = a_0$ .

In  $a(t, n)$  secret-sharing system,  $n$  points are randomly chosen as  $x_i : 1 \leq i \leq n$ , and  $x_i \neq 0 \in Z_p$ , the

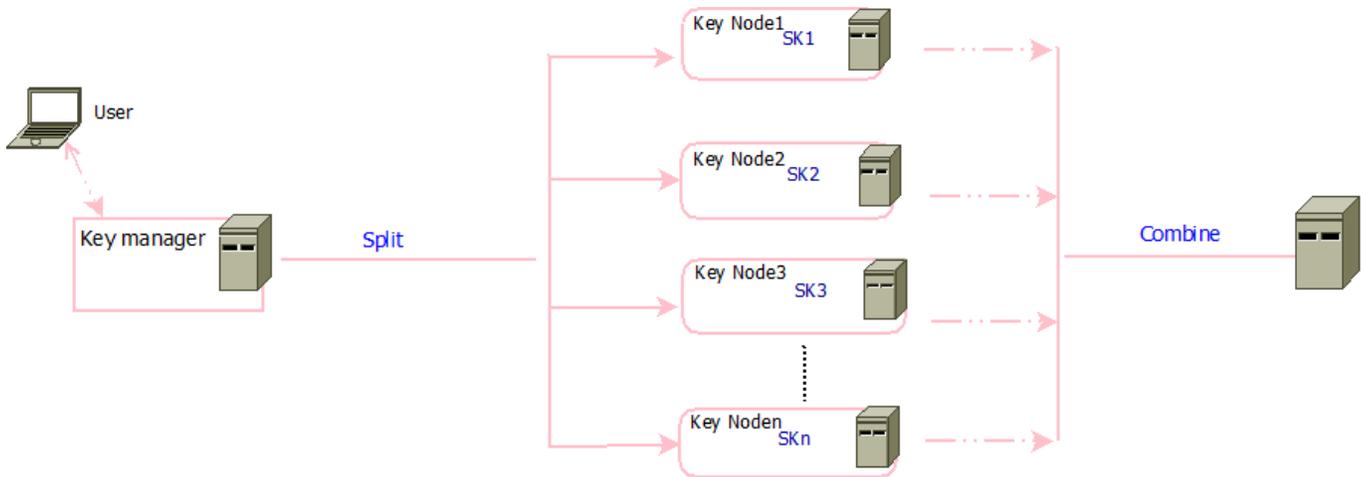


Fig. 5. Architecture AuSDiDe.

dealer calculates  $y_i = f(x)$  and transmits  $s_i = (x_i, y_i)$  to shareholders  $U_i$ .

For a secret reconstruction [18], supposing that  $t$  shareholders  $U_1, U_2, \dots, U_t$ . Each shareholder  $U_i$  gives a share  $s_i$  to the other stockholders. After that, if a shareholder possesses  $m$  shares  $s_1, s_2, \dots, s_m$ , he may retrieve  $f'(x)$  using the Lagrange interpolation polynomial as [21]:

$$f'(x) = \sum_{i=1}^t s_i \prod_{j=1, i \neq j}^t \frac{x_j - x}{x_j - x_i}, (mod p) \quad (2)$$

The secret  $S$  will be computed as:

$$f'(0) = \sum_{i=1}^t s_i \prod_{j=1, i \neq j}^t \frac{x_j}{x_j - x_i}, (mod p) \quad (3)$$

#### IV. AUSDiDE GENERAL ARCHITECTURE

The AuSDiDE is introduced to enhance security and been developed to increase the level of security. This system takes into consideration those architectures: distributed and centralized.

AuSDiDe is composed of nodes playing different exclusive roles between them. Hacking and attacking servers should be difficult. Getting a key will not be enough to encrypt the passphrase, you need to get all shared keys. It will be complicated, given the number of servers.

The AuSDiDe is composed of two main Node as shown in Fig. 4: Key Manager – the master and Key Node – the slave. One of the machines is the master, called Key Manager: This machine contains all names and key parts, like a phone book.

All other machines are Key Node. They store the different shared secret key. The key Manager knows where the keys are, which part of the key and on which key Manager are registered.

The key manager will be dividing the secret into several parts –  $(n, k)$  key parts after having defined the threshold  $k$ . As illustrated in Fig. 5, this task called split operation.

The threshold represents the minimum number of parties necessary to reconstitute the passphrase and unlock access to the secret. This task called combine operation.

#### V. IMPLEMENTATION

In this section, the implementation of the AuSDiDe is presented. The created cluster is composed of different servers, playing different exclusive roles. For this operation, the docker [22] and Hadoop framework [23] are used.

##### A. Docker Security Advantages

Docker’s most widely used containerization technique can raise the degree of security if used correctly (in comparison to running applications directly on the host). On the other hand, misconfigurations might result in a reduction in security or even the introduction of new vulnerabilities. Docker gives the ability to automate the deployment of applications into Containers [24]. Docker offers an additional layer of deployment engine on top of a Container environment where programs are virtualized and executed. Docker is meant to offer a speedy and lightweight environment in which code may be executed quickly and an additional facility of the competent work process to remove the code from the computer for testing before production [24]. You may certainly start with a docker with a basic configuration system, a docker binary with a Linux kernel. Docker has four major internal components: Docker Server and Client, Docker Registries, Docker Images, and Containers. A Docker image is used to generate a Docker container [22]. Containers store all of the components needed for a program, allowing it to execute in isolation. Assume there is an image of Ubuntu OS with MongoDB server; when executed with the docker run command, a container is produced, and MongoDB server is operating on Ubuntu OS [22].

##### B. Benefits and Advantages of Hadoop

The Apache Hadoop project creates open-source software for scalable, distributed computing. The software library is a framework that enables the distributed processing of massive

data volumes across computer clusters by using basic programming techniques [23]. It is intended to grow from a single server to thousands of computers, providing local computing and storage.

Rather than relying on hardware to provide high availability, the library is intended to identify and manage failures at the application layer, allowing a highly available service to be delivered on top of a cluster of machines that may all fail [23]. Therefore, it is widely used today to store, analyze, and manipulate huge amounts of data: Hadoop is a standard for Big Data processing. Hadoop refers to its ecosystem and all software such as Apache Spark, Cloudera Impala, Sqoop, etc.

The Experimental Architecture of AuSDiDe Implementation is illustrated in Fig. 6.

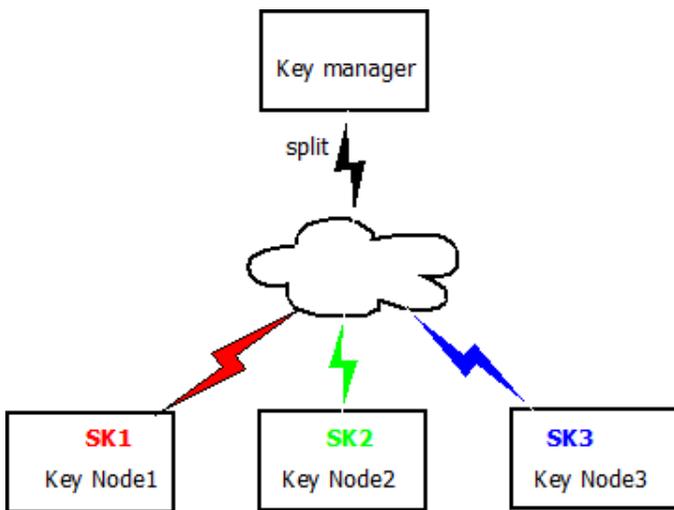


Fig. 6. Experimental Architecture of AuSDiDe Implementation.

The three containers represent a Key Manager and three Key nodes utilizing a Docker image. We use throughout the implementation three containers, representing respectively a Key Manager and three Key Nodes using docker image [24], as shown in Fig. 6.

The result of this execution shows how the split operation at the Key Manager was able to divide and share the shared keys SK1, SK2 and SK3 and store them to the different Key Node1, Key Node2 and Key Node3.

For the inverse operation i.e combine operation, the threshold equal two, representing the minimum number of parties necessary to reconstitute the passphrase and unlock access to the secret. It can be SK1 and SK2; SK2 and SK3 or SK1 and SK3.

## VI. DISCUSSION

Several considerations must be taken to determine the authentication system adapted for new architectures. The future of internet (for example web 3.0) implement a decentralized architecture, all machines can play the role of master and slave at the same time. For example, Peer-to-peer (P2P) is a model of computer network structured in a decentralized way [25], the communications between nodes with equal responsibility. In

the world network, nodes are identified by a logical IP address. To maintain anonymity in a network, it is better to use private or public key addresses. These different constraints have been well studied for the development of the AusDiDe solution.

## VII. CONCLUSION

This paper proposes a new approach focused on authentication systems, private keys, and password management. This solution presents a new way of thinking about authentication systems and sensitive data security. These will be adapted to future generation Internet, e.g., web3. The implementation of AuSDiDe clearly shows better security and an obstacle for malicious attacks. The AuSDiDe system divides and shares keys toward different and distributed nodes in all clusters. Hacking requires obtaining all the shared keys, making it difficult for hackers. As continuity to this work and to enhance the AuSDiDe functionality, the artificial intelligence concept will develop an intelligent AuSDiDe system operating in decentralized and hybrid architecture as Blockchain. Towards a framework for a smart AuSDiDe with a learning system to anticipate intrusions and anomalies in order to reinforce security.

## REFERENCES

- [1] BERDIK, David, OTOUM, Safa, SCHMIDT, Nikolas, et al. A survey on blockchain for information systems management and security. *Information Processing and Management*, 2021, vol. 58, no 1, p. 102397.
- [2] YANG, Caixia, TAN, Liang, SHI, Na, et al. AuthPrivacyChain: A blockchain-based access control framework with privacy protection in cloud. *IEEE Access*, 2020, vol. 8, p. 70604-70615.
- [3] SHARMA, Uttam, TOMAR, Pradeep, ALI, Syed Sadaf, et al. Optimized Authentication System with High Security and Privacy. *Electronics*, 2021, vol. 10, no 4, p. 458.
- [4] JEONG, Junho, KIM, Donghyo, IHM, Sun-Young, et al. Multilateral Personal Portfolio Authentication System Based on Hyperledger Fabric. *ACM Transactions on Internet Technology (TOIT)*, 2021, vol. 21, no 1, p. 1-17.
- [5] KRISHNARAJ, N., BELLAM, Kiranmai, SIVAKUMAR, B., et al. The Future of Cloud Computing: Blockchain-Based Decentralized Cloud/Fog Solutions—Challenges, Opportunities, and Standards. In : *Blockchain Security in Cloud Computing*. Springer, Cham, 2022. p. 207-226.
- [6] GAI, Keke, GUO, Jinnan, ZHU, Liehuang, et al. Blockchain meets cloud computing: A survey. *IEEE Communications Surveys and Tutorials*, 2020, vol. 22, no 3, p. 2009-2030.
- [7] VERGNE, J. P. Decentralized vs. distributed organization: Blockchain, machine learning and the future of the digital platform. *Organization Theory*, 2020, vol. 1, no 4, p. 2631787720977052.
- [8] ZARRIN, Javad, PHANG, Hao Wen, SAHEER, Lakshmi Babu, et al. Blockchain for decentralization of internet: prospects, trends, and challenges. *Cluster Computing*, 2021, p. 1-26.
- [9] MASLOUHI, Imane, GHOUMID, Kamal, ZAIDOUNI, Jamal, et al. Network Higher Security of Intelligent Networks Platforms in Telecommunications Areas. In : *Proceedings of the Mediterranean Conference on Information and Communication Technologies 2015*. Springer, Cham, 2016. p. 225-233.
- [10] FATIMA, Shahin et AHMAD, Shish. Secure and effective key management using secret sharing schemes in cloud computing. *International Journal of e-Collaboration (IJeC)*, 2020, vol. 16, no 1, p. 1-15.
- [11] FĂLĂMAȘ, Diana-Elena, MARTON, Kinga, et SUCIU, Alin. Assessment of Two Privacy Preserving Authentication Methods Using Secure Multiparty Computation Based on Secret Sharing. *Symmetry*, 2021, vol. 13, no 5, p. 894.
- [12] GUPTA, Kishor Datta, RAHMAN, Md Lutfar, DASGUPTA, Dipankar, et al. Shamir's Secret Sharing for Authentication without Reconstructing Password. In : *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 2020. p. 0958-0963.

- [13] HENA, M. et JEYANTHI, N. Blockchain Based Authentication Framework for Kerberos Enabled Hadoop Clusters. In : *Soft Computing for Problem Solving*. Springer, Singapore, 2021. p. 315-327.
- [14] BISSOLI, Andrea et D'AMORE, Fabrizio. Authentication as a service: Shamir Secret Sharing with byzantine components. *arXiv preprint arXiv:1806.07291*, 2018.
- [15] JAROUCHEH, Zakwan et ÁLVAREZ, Iván Abellán. Secretation: Toward a Decentralised Identity and Verifiable Credentials Based Scalable and Decentralised Secret Management Solution. In : *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. IEEE, 2021. p. 1-9.
- [16] SATVAT, Kiavash, SHIRVANIAN, Maliheh, et SAXENA, Nitesh. PAS-SAT: Single Password Authenticated Secret-Shared Intrusion-Tolerant Storage with Server Transparency. *arXiv preprint arXiv:2102.13607*, 2021.
- [17] ESPOSITO, Christian, FICCO, Massimo, et GUPTA, Brij Bhooshan. Blockchain-based authentication and authorization for smart city applications. *Information Processing and Management*, 2021, vol. 58, no 2, p. 102468.
- [18] DZURENDA, Petr, RICCI, Sara, MARQUÉS, Raúl Casanova, et al. Secret Sharing-based Authenticated Key Agreement Protocol. In : *The 16th International Conference on Availability, Reliability and Security*. 2021. p. 1-10.
- [19] BHATTACHARJEE, Arpan, BADSHA, Shahriar, SHAHID, Abdur R., et al. Block-phasor: A decentralized blockchain framework to enhance security of synchrophasor. In : *2020 IEEE Kansas Power and Energy Conference (KPEC)*. IEEE, 2020. p. 1-6.
- [20] SAROSH, Parsa, PARAH, Shabir A., et BHAT, Ghulam Mohiuddin. Utilization of secret sharing technology for secure communication: a state-of-the-art review. *Multimedia Tools and Applications*, 2021, vol. 80, no 1, p. 517-541.
- [21] LI, Guojia et YOU, Lin. A Consortium Blockchain Wallet Scheme Based on Dual-Threshold Key Sharing. *Symmetry*, 2021, vol. 13, no 8, p. 1444.
- [22] RAD, Babak Bashari, BHATTI, Harrison John, et AHMADI, Mohammad. An introduction to docker and analysis of its performance. *International Journal of Computer Science and Network Security (IJCSNS)*, 2017, vol. 17, no 3, p. 228.
- [23] The Apache Hadoop project.URL: <https://hadoop.apache.org/>. access on Dec. 2021
- [24] AHMED SHAIKH, Kasam et AGASKAR, Shailesh S. Containers and Azure Kubernetes Services. In : *Azure Kubernetes Services with Microservices*. Apress, Berkeley, CA, 2022. p. 103-129.
- [25] KUSHWAHA, Satpal Singh, JOSHI, Sandeep, SINGH, Dilbag, et al. Systematic Review of Security Vulnerabilities in Ethereum Blockchain Smart Contract. *IEEE Access*, 2022.

# Keyphrases Concentrated Area Identification from Academic Articles as Feature of Keyphrase Extraction: A New Unsupervised Approach

Mohammad Badrul Alam Miah<sup>1</sup>

Faculty of Computing

Universiti Malaysia Pahang, Pekan, Malaysia

Information and Communication Technology

Mawlana Bhashani Science and Technology University,

Tangail, Bangladesh

Suryanti Awang<sup>2</sup>

Faculty of Computing

Centre for Data Science & Artificial

Intelligence (Data Science Centre)

Soft Computing & Intelligent Systems

Universiti Malaysia Pahang, Pekan, Malaysia

Md. Saiful Azad<sup>3</sup>

Computer Science and Engineering

Green University of Bangladesh

Dhaka, Bangladesh

Md Mustafizur Rahman<sup>4</sup>

Department of Mechanical Engineering

Faculty of Engineering

Universiti Malaysia Pahang, Gambang, Kuantan, Malaysia

**Abstract**—The extraction of high-quality keywords and summarising documents at a high level has become more difficult in current research due to technological advancements and the exponential expansion of textual data and digital sources. Extracting high-quality keywords and summarising the documents at a high-level need to use features for the keyphrase extraction, becoming more popular. A new unsupervised keyphrase concentrated area (KCA) identification approach is proposed in this study as a feature of keyphrase extraction: corpus, domain and language independent; document length-free; utilized by both supervised and unsupervised techniques. In the proposed system, there are three phases: data pre-processing, data processing, and KCA identification. The system employs various text pre-processing methods before transferring the acquired datasets to the data processing step. The pre-processed data is subsequently used during the data processing step. The statistical approaches, curve plotting, and curve fitting technique are applied in the KCA identification step. The proposed system is then tested and evaluated using benchmark datasets collected from various sources. To demonstrate our proposed approach's effectiveness, merits, and significance, we compared it with other proposed techniques. The experimental results on eleven (11) datasets show that the proposed approach effectively recognizes the KCA from articles as well as significantly enhances the current keyphrase extraction methods based on various text sizes, languages, and domains.

**Keywords**—Keyphrase concentrated area; KCA identification; feature extraction; data processing; keyphrase extraction; curve fitting

## I. INTRODUCTION

The continuous development of the information age and exponential growth of textual information makes it even more challenging to handle this large amount of information [1]. Before the emergence of technology, this information could be processed by humans, which was very time-consuming. Furthermore, due to the inconsistencies between the amount of data and manual data processing skills, it is challenging to

complete this vast information, leading to automated keyphrase extraction systems that utilise computers' extensive computational capability to substitute manual labour [2], [3].

The goal of automated keyword/keyphrase extraction techniques is to extract high-quality keys from documents. In general, Keyphrase offers a high level of description, summary, and characterization of documents, which is crucial for many aspects of Natural Language Processing, such as articles categorization, classification, and clustering [3]. They are, nevertheless, used in a wide range of Digital Information Processing applications, including Digital Content Management, Information Retrieval [3], [4], Contextual Advertising [5], and Recommender System [6]. It also offers a wide range of practical uses, including media searches, search engines, digital libraries, legal and geographic information retrieval [7].

Various keyphrase extraction methods have been developed to support the aforementioned applications [8], [9], [7], [10], [11], [12]. Domain-specific strategies [9], for example, need knowledge of the application domain, whereas linguistic approaches [9] demand language proficiency. They cannot solve problems in other disciplines or languages as a result. Supervised techniques need a lot of unusual train data to extract the quality keyphrases. Owing to their vast number of complicated operations, unsupervised machine learning methods are computationally costly, and they perform badly due to their inability to identify cohesiveness among several words that make up a keyword [7], [13], [14], [15]. Feature extraction is essential for those keyphrase extraction methods that want high-quality keyphrases. It's the process of obtaining characteristics (sometimes referred to as features) that distinguish keywords from other terms [16]. These features also impact the performance of various supervised and un-supervised keyword/keyphrase extraction methods. It is demonstrated that from the previous debate, the feature extraction of keyphrases remains an essential research topic for the study.

Therefore, this article proposes an unsupervised new Keyphrase Concentrated Area identification technique with ensuing significant contributions:

- The proposed technique, which is corpus-independent, can be applied to any text and any corpus.
- KCA identification's a domain- and language-agnostic method that relies on little statistical knowledge.
- The proposed method can be used as a keyphrase feature in both supervised and unsupervised approaches.
- It's a document length-free refers to the fact that there are no requirements for the minimum length of a document that a keyphrase must-have.
- Eleven datasets have been used to test and assess the effectiveness of the proposed method.

The remainder of this paper is organised as follows. Section II outlines the various methodologies, including their benefits and drawbacks, and so emphasises the need for a new strategy to be proposed. The suggested technique is then discussed in depth in Section III. The setup of the experiments is detailed in Section IV, which contains corpus data, evaluation measures, and implementation details. In Section V, all of the obtained findings are plotted and analysed, and Section VI brings this article to a close.

## II. RELATED WORK

This section will discuss similar strategies because the proposed technique is a novel approach for extracting keyphrase features. Most keyphrase extraction techniques are categorized into two groups such as supervised and unsupervised, based on the training datasets [4]. Feature extraction is used in both ways. Below, we'll go over the main points of both of these groups' approaches.

### A. Supervised Methods

The keyphrase extraction technique is counted as a binary classification problem [1] using this method from articles, with a proportion of candidate keyphrases categorised as keyphrases and non-keyphrase. Methods for solving the classification problem include support vector machines, Decision trees, Naive Bayes [3], Neural networks [17], [18], and C4.5 [19]. The prominent techniques are examined in detail in the subsequence that adopts this method.

As a feature, Key Extraction Algorithm (KEA) [20] uses TFxIDF and the first presence location. It utilises descriptive approaches for identifying candidate keyphrases, estimating feature values for each candidate and predicting and determining candidates' good keyphrases using the Naive Bayes algorithm. However, KEA depends on the training dataset, and if the training dataset does not match the documents, it may produce poor results.

As a feature, Genitor Extractor (GenEx) [1] assigns first occurrence position, term frequency (TF), and keyphrase length. The most well-known key extraction approach is established on a collection of parametrized heuristic rules that employ genetic algorithms to retain their efficacy across diverse domains, and it is based on a C-4.5 decision-making process. It does not use

the Term Frequency-Inverse Document Frequency technique (TF-IDF).

Unlike the GenEx and KEA methods, the Hult system [1] allows the extracted keys to be as long as they want to be. The four characteristics it utilises are part of speech (POS) tag, n-grams, noun phrase (NP) chunks, first occurrence position, and TF. Unfortunately, no association exists between the various POS tag features. The system doesn't test on KEA or GenEx corpus, and the stated recall value is poor.

The Maui Algorithm [21], based on the KEA system, is an automatic generic topical indexing algorithm. It adds data from Wikipedia to expand the KEA system. However, one of this algorithm's flaws is its lack of assessment abilities.

The position of a term, its first occurrence; phrases; informativeness; keywords; and the length of the candidate term as a feature are all used by HUMB [22]. In a variety of data sets, the HUMB system has produced positive results. HUMB, on the other hand, has only used scientific papers.

The Document Phrase Maximality (DPM)-index, first position, TF, TFxIDF, IDF, first sentence, average sentence length, head frequency, substrings frequencies sum, and five other new features are (18 statistical features) used by DPM-index [23]. Without external knowledge or document structural elements, this system's results have improved significantly compared to other keyphrase extraction systems.

Citation-enhanced Keyphrase Extraction (CeKE) [24] utilize the following keyphrase features such as TFxIDF, relative Pos, inCited, POS, first position, inCiting, TF-IDF-Over, firstPosUnder, citation TF-IDF. They can improve keyphrase extraction and add keyphrase features. (CeKE + keys) the model outperforms other systems [1].

Keyphrase Extraction (KeyEx) Method [25] finds a large number of possible candidate keyphrases and build a classification model for key extraction using supervised learning methods. Experiments conducted by the author revealed that the KeyEx system has effectively improved the extracted keyphrase's quality. In addition, their strategy beats existing sequential pattern mining methods.

### B. Unsupervised Methods

The keyphrase extraction scheme is a ranking issue that is solved without prior knowledge. These methods can be classified as statistical or graph-based [1]. The following sections go over the most important techniques used by both groups in sufficient detail.

PageRank [26] is a graph-based algorithm that uses random walks as its foundation. It is, however, appropriate for raking web and social media pages but not for extracting keyphrase from formal documents. PageRank extension known as PositionRank [14] was discovered to improve performance, which scores word by taking into account all of its positions and its frequency, and thus determines its rank. This technique, however, poorly performs because it ignores topical coverage and diversity.

TextRank [27] uses Parts of Speech (POS) as an internal feature, with several limitations, including the inability to

capture cohesiveness, resulting in sub-optimal results. TopicRank [28] is another keyphrase extraction technique that overcomes TextRank's limitations. The noun phrases in the document are extracted and clustered into topics by TopicRank. Furthermore, it has an issue with error propagation. The lengthening of TextRank is SingleRank [29]. It correctly pulls only noun phrases from the records, not keyphrases, by collecting ranked words. However, it does not always filter out low-scoring words and gives longer keys higher scores, but non-significant keys are included in the ranking process.

MultipartiteRank [15] is a technique for resolving the TopicRank error propagation problem. However, it suffers from clustering error, making selecting the most representative candidates challenging. Tree-based Keyphrase Extraction Technique (TeKET) [7] is a renowned unsupervised keyphrase extraction method that is language and domain-independent and needs only rudimentary statistical knowledge. Though it outperforms some other keyphrase extraction techniques, it has some disadvantages, such as tremendous flexibility.

The most common statistical method is named TF-IDF [30]. Although TF-IDF is simple to implement, computing Inverse Document Frequency (IDF) takes a long time and requires a lot of computing power when dealing with a large dataset. The KP-Miner [31] program is used to solve the problem of single-term preference. Although KP-Miner exceeds TF-IDF, it still has some drawbacks, including degrading the global ranking performance if the number of records increases. It's also computationally expensive because it relies on TF-IDF.

Yet Another Keyword Extractor (YAKE) [10] is another popular technique for removing the IDF problem by calculating the weighting score of a keyphrase using five features/attributes: as term position, casing, term relatedness to context, term frequency normalization, and term distinct sentence. However, because it uses the N-grams technique to generate candidate keys, its computational complexity grows linearly with N-grams.

According to the previous discussions, both supervised and unsupervised keyphrase extraction techniques have several drawbacks that prevent them from achieving better results. Therefore, this paper proposes a new unsupervised KCA identification technique as a keyphrases feature that will significantly decrease the specified flaws as well as extract high-quality keywords from academic articles.

### III. METHODOLOGY

The whole approach of keyphrase concentrated area identification utilizing the proposed method is divided into three major stages: *i*) Data preprocessing, *ii*) Data processing, and *iii*) KCA identification (see Fig. 1). In the subsequent sections, the proposed strategy is illustrated in detail.

#### A. Data Pre-processing

It is an important stage in the development of our proposed technique. Initially, the proposed approach gathered eleven datasets (having 9006 papers) covering three languages (Portuguese, English, and Spanish), different disciplines (such as chemistry, physics, computer science, and others). Containing

four different kinds of papers (news, abstracts, full articles, and M.Sc/Ph.D. Thesis) ranging from 75 tokens to 8000 tokens per document) [32]. Every dataset has two kinds of file names, like keys and docsutf8, including the same articles/documents. Visit Section IV-A for more information.

After that, the suggested method extracts the docsutf8 files (which include various vital articles as text files) as well as the keys files independently (containing different essential keys known as text files). Afterward, read these two files and save them respectively as document ( $\delta$ ) and keys ( $\chi$ ). After receiving the documents and keys, they must normalize the data, which entails four steps: Convert the document to lower case; Eliminate the irrelevant numbers by employing regular expressions); Remove all punctuation marks; Remove blank spaces (using the strip() function to remove leading and to end spaces) [33]. After that, The splitting technique is applied on keys files to compute the keyphrase learned as GoldKey ( $\gamma$ ) founded on Newline ( $\backslash n$ ) method. At that moment, in our proposed approach, the length of text or document is split into ten (10) and twenty (20) regions.

#### B. Data Processing

This is a crucial step after pre-processing the data. During this step, the proposed system uses the first appearance to locate (*Loc*) of each ( $\gamma$ ) of ( $\chi$ ) from the ( $\delta$ ). Save the *Loc* of  $\gamma$  in the proper region of the  $\delta$  if located in the  $\delta$ . Note that the *Loc* is stored on two-dimensional (2D) array in which column is the ( $\delta$ ) region's number and row is the ( $\gamma$ )'s number. If the  $\gamma$  is not located, research the  $\delta$  for the next  $\gamma$  of  $\chi$ . This procedure will repeat until  $\gamma$  has completed the  $\chi$  file for a single dataset document. The same procedure will continue for all datasets.

#### C. KCA Identification

It is an important and final phase after data processing. The output of the data processing phase is applied to this phase to find the concentration area of the keyphrases. This phase consists of the three significant steps: *i*) Average value calculation, *ii*) Curve plotting, and *iii*) Curve fitting technique that describes the following sections.

*a) Average Value Calculation:* To begin, for a single document/text, compute the Average (Avg) value of every region and save it in a new 2D array whose row is the number of records in a particular dataset and column is the text/document regions like as before. Afterwards, the process will resume until every document for a specific dataset has been completed. Calculate the average value of every region/portion for every record in a particular dataset and save this average value in another new 2D array whose row is the entire dataset and column is the same as before. After that, the Avg calculation will resume until every dataset has been completed [3]. Definitely, for all datasets, compute the Avg value of all regions.

*b) Curve Plotting (CP):* CP is a graphical presentation approach for a dataset. It's possible to read plotted values as known functions of unknown variables using this method. In data analysis and statistics it is pretty useful. CP is used to understand our proposed method's keyphrases concentration region/area. Because of this, the Avg value of each dataset is plotted alongside the Avg value of the whole dataset.

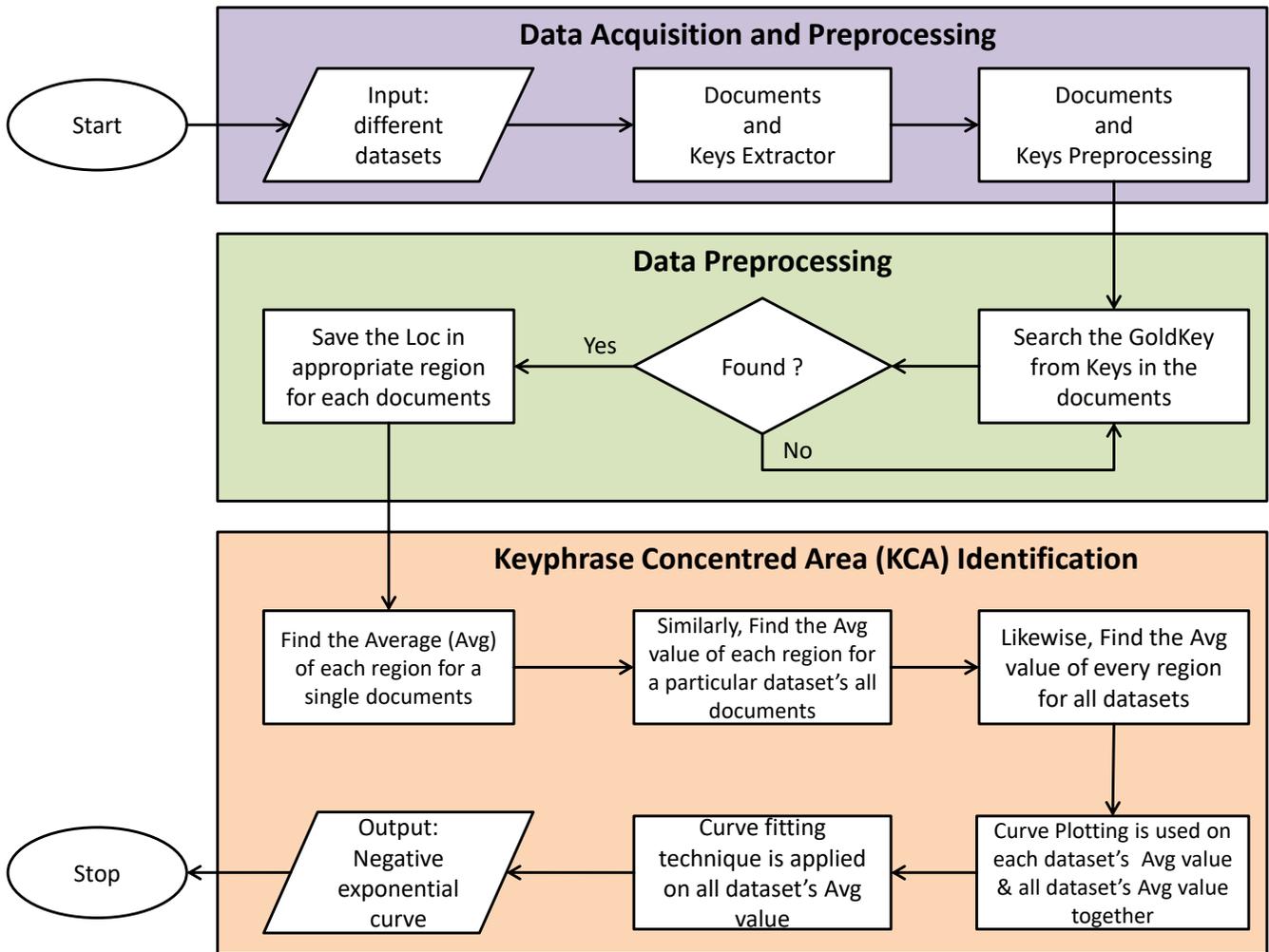


Fig. 1. The Proposed Technique's Flow Diagram for KCA Identification.

c) *Curve Fitting Technique (CFT)*: It is a helpful method for analysing linear, polynomial, and nonlinear curves. It is most likely the process of producing the best-fitting curve or mathematical function for a constrained set of data points. CFT is used to identify the critical concentration region/area in our proposed approach. As a result, CFT is applied on the Avg value of all datasets, resulting in a negative exponential curve for the proposed approach.

#### IV. EXPERIMENTAL SETUP

Our proposed method clearly stated that the experimental setting introduces corpus/dataset details, implementation details, and evaluation metrics, presented in the following section. Afterwards, the outcomes are explained in Section V.

##### A. Corpus Details

our proposed approach has tested on 11 datasets/corpus to evaluate the performance. How the proposed approach behaves under many datasets was our another ambition to understand. Standard gatherings such as Inspec [32], SemEval2010 [34], 110-PT-BN-KP [35], Nguyen2007 [36],

PubMed [32], Schutz2008 [37], catic [38], kdd [39], wicc [38], www [39], and theses100 [32] are used in our proposed approach. A quick summary is given in the preceding section III-A, and a statistical review of all datasets is given in Table I. Every corpus is explained in detail in the following sections.

**Inspec** [32] contains 2000 abstracts and 28220 gold keys from computer science articles published from 1998 to 2002. There are two sets of keywords in each document: controlled keywords, selected manually from the Inspec vocabulary, and uncontrolled keywords, which the editors liberally allocate.

**WWW** [39] and **KDD** [39] are the tiniest datasets (on an Avg of 84 and 75 tokens per document). The collection of those datasets (like Inspec) is based on abstracts of papers published between 2004 and 2014 at the ACM Conference and the World Wide Web(WWW) Conference on Knowledge Discovery in Databases (KDD). There are 1,330 and 755 documents in each and 6405 and 3093 goldkeys.

**SemEval2010** [34] is one of the famous standard datasets, which contains 244 whole scientific articles extracted from the ACM Library. The papers range in length from 6 to 8 pages

TABLE I. A STATISTICAL DATASETS SUMMARY FOR THE ANALYSIS OF PRESENT AND ABSENT GOLDKEYS

| Dataset      | Language | Type of Doc    | Domain        | #Docs | #Gold Keys | #Present Goldkey | #Absent Goldkey | Absent Goldkey per doc(%) | Present Goldkey per doc(%) |
|--------------|----------|----------------|---------------|-------|------------|------------------|-----------------|---------------------------|----------------------------|
| 110-PT-BN-KP | PT       | News           | Misc.         | 110   | 2688       | 2616             | 72              | 1.34%                     | 98.66%                     |
| Cacic        | ES       | Paper          | Comp. Science | 888   | 3396       | 3057             | 339             | 10.44%                    | 89.56%                     |
| Inspec       | EN       | Abstract       | Comp. Science | 2000  | 28220      | 12007            | 16213           | 55.98%                    | 44.02%                     |
| Kdd          | EN       | Paper          | Comp. Science | 755   | 3093       | 1031             | 2062            | 65.78%                    | 34.22%                     |
| Nguyen2007   | EN       | Paper          | Comp. Science | 209   | 2507       | 2008             | 499             | 18.96%                    | 81.04%                     |
| PubMed       | EN       | Paper          | Comp. Science | 500   | 7120       | 2513             | 4607            | 63.91%                    | 36.09%                     |
| Schutz2008   | EN       | Paper          | Comp. Science | 1231  | 55718      | 47387            | 8331            | 14.79%                    | 85.21%                     |
| SemEval2010  | EN       | Paper          | Comp. Science | 243   | 3785       | 3129             | 656             | 17.12%                    | 82.88%                     |
| Theses100    | EN       | MSc/PhD Thesis | Misc.         | 100   | 667        | 302              | 365             | 55.14%                    | 44.86%                     |
| Wicc         | ES       | Paper          | Comp. Science | 1640  | 5860       | 5275             | 585             | 9.16%                     | 90.84%                     |
| WWW          | EN       | Paper          | Comp. Science | 1330  | 6405       | 2122             | 4283            | 64.68%                    | 35.32%                     |

and cover four distinct areas of computer science: information search and retrieval, Distributed artificial intelligence, Distributed Systems, and Social and behavioural sciences. Every paper has a set of keyphrases assigned by the author as well as by professional editors.

**Nguyen2007** [36]: There are 209 scientific conference papers and 2507 gold keys in this dataset. Three articles were provided to student volunteers to read, and the goldkeys were handed out manually. Each document has twelve(12) goldkeys on Avg.

Both **Schutz2008** [37] and **PubMed** [32] are corpuses compiled from a PubMed Central full-text paper that cites over 26 million online books of life science journals from MIDLINE. Schutz2008 is made up of 1,231 articles chosen from PubMed Central, whereas PubMed is made up of 500 articles chosen from identical sources. The authors' Schutz2008 keyword is hidden in the paper and employed as goldkeys, yielding 45.26 goldkeys per document. The gold keyword in PubMed is Medical Subject Headings (MeSH), which is a controlled vocabulary glossary utilised to index articles, occurring in 14.24 goldkeys in each document.

**Theses100** [32] corpus comprises of hundred(100) complete Masters and PhD thesis from University of Waikato, New Zealand. These domains are relatively dissimilar, departing from computer science, chemistry, economics, philosophy, psychology, history, etc. It has 6.67 goldkeys per document, on Avg.

**110-PT-BN-KP** [35] is a Television(TV) Broadcast News(BN) corpus including 110 transcripts from eight(8) broadcast news programmes from the European Portuguese ALERT BN corpus, including finance, sports, politics, and others theme. Goldkeys were created by having a tagger remove all keywords that contained document content summaries, yielding 24.44 goldkeys per document.

**Cacic** [38] consists of 888 scientific publications from 2005 to 2013. It also comprises the minor number 3.82 goldkeys in each document, on Avg. The **Wicc** [38] dataset made up of

TABLE II. CONFUSION MATRIX

|                          |  | Actual Class Positive | Actual Class Negative |
|--------------------------|--|-----------------------|-----------------------|
| Predicted Class Positive |  | $T_P$                 | $F_N$                 |
| Predicted Class Negative |  | $F_P$                 | $T_N$                 |

1,640 scientific papers published from 1999 to 2012, with an Avg of 3.57 goldkeys in each document.

### B. Evaluation Metrics

*Accuracy, error rate, recall, precision, F<sub>1</sub>-score*, and other significant and relevant metrics are routinely used to measure the performance of a system. To evaluate the performance of our proposed approach, we employ accuracy data and a confusion matrix (shown in Table II). The accuracy measure is generally defined as the percentage of correct predictions out of the total number of patterns analysed. The following equation (1) represents *accuracy*.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \quad (1)$$

Here, True Positive ( $T_P$ ) and True Negative ( $T_N$ ) denote the number of positive and negative keyphrases accurately classified, respectively. On the other hand, False Positive ( $F_P$ ) and False Negative ( $F_N$ ) represent the number of positive and negative keywords that were wrongly classified.

### C. Implementation Details

Python 3.6 and the Spyder-IDE are used to implement the proposed method. It is a high-level and object-oriented programming language that is easy to learn and utilise. It has a data structure that is user-friendly, versatile, and supported by numerous libraries. It increases productivity, is interpreted, dynamically typed, and is free and open-source. It is applied in big data, Cloud Computing, and Machine Learning, etc.

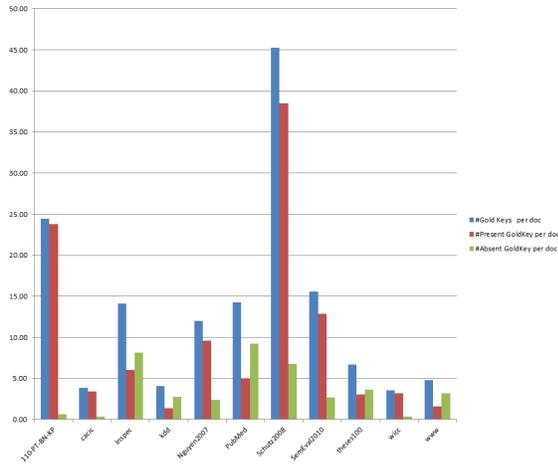


Fig. 2. The Avg Number of Goldkeys are Present and Absent in Each Document for All Datasets.

Following that, the machine is outfitted with an Intel Core i7 processor, RAM-12GB, a SATA-connected solid state drive (SSD), and the Windows 10 operating system [3].

## V. RESULTS AND DISCUSSION

This section includes a full examination of the experiment outcomes. The proposed system divides the text or documents length into twenty (20) and ten (10) regions to identify the Keyphrases Concentrated Area (KCA). When more than twenty regions are raised, the first region produces significantly less goldkey than twenty regions. Similarly, if the number of regions is lowered to less than 10, the first region has significantly more goldkey than ten regions. Our proposed technique aims to locate the KCA in documents/articles; thus, instead of expanding or lowering the regions, the system is examined for all types of text lengths as ten and twenty regions. This section is divided into two phases described in the following section: *i)* Result Analyses, and *ii)* Comparison of Proposed Systems.

### A. Results Analysis

The proposed system's performance is evaluated in this phase using the following criteria: *i)* Dataset Analysis, *ii)* Plotting Analysis, and *iii)* Curve Fitting Analysis, are the three types of results analysis.

*a) Dataset Analysis:* The proposed system has been tested on eleven (11) datasets (detail in section IV-A) to judge the performance of the proposed technique. Afterwards, the proposed system determines how many documents, number of goldkeys, present and absent goldkeys, as well as present and absent goldkeys in each article in (%) exist in every dataset provided in Table I based on the analysis of the datasets. The Avg number of goldkeys present and absent per document are examined for each dataset, exhibited in Fig. 2. Likewise, the Avg number of goldkeys absent and present in percentage(%) of each document for all datasets is displayed in Fig. 3. According to our findings, 65.70% of goldkeys per document are present on Avg across all datasets, while 34.30% are absent.

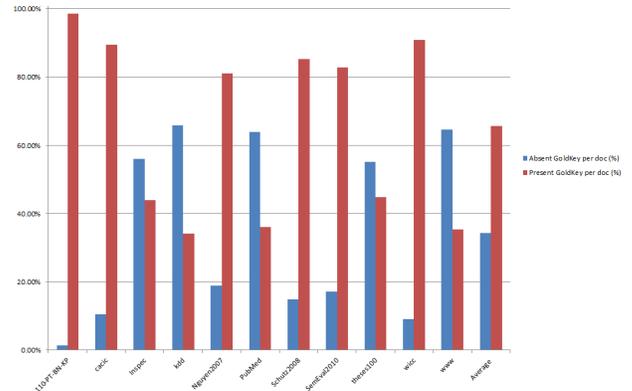


Fig. 3. The Avg Number of Goldkeys are Absent and Present in Percentage per Document for all Datasets.

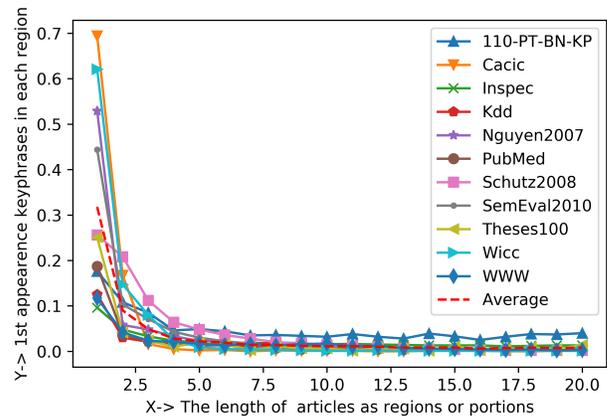


Fig. 4. The Plotting Analysis of KCA Identification by Considering 1st Appearance Keyphrases for 20 Regions.

*b) Plotting Analysis:* According to the previous discussion, Since the Avg of 65.70% of goldkeys is present per document for each dataset, all the results in this work have been predicated on 65.70% of present goldkeys. The first appearance keyphrases in a document are considered in our proposed method, and the text length is divided into twenty(20) and ten(10) regions. The proposed method then plots the eleven (11) dataset's values and Avg value of all datasets together based on each region of articles. Fig. 4 shows the analysis of first appearance keyphrases in each region for KCA identification when the text length is divided into twenty(20) regions. Similarly, Fig. 5 shows the analysis of first appearance keyphrases in each region for KCA identification when the text length is divided into ten(10) areas/regions. Since all dataset curves together are negative exponential, it is confirmed that the maximum goldkeys/keyphrases are found in 1st region, then 2nd region of the articles, and so forth, as shown in Fig. 4 and Fig. 5.

*c) Curve Fitting Analysis:* After completing the plotting analysis, the Avg value of entire datasets is applied in this analysis of our proposed system. Afterwards, the system attempts to discover the first fitted curve and then the negative exponential equation for each region's Avg value. In Fig. 6, the

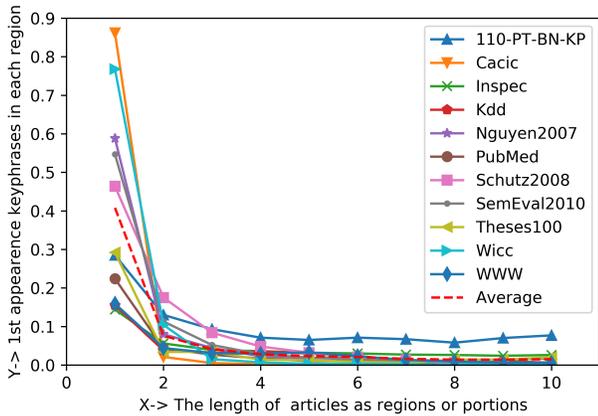


Fig. 5. The Plotting Analysis of KCA Identification by Considering 1st Appearance Keyphrases for 10 Regions.

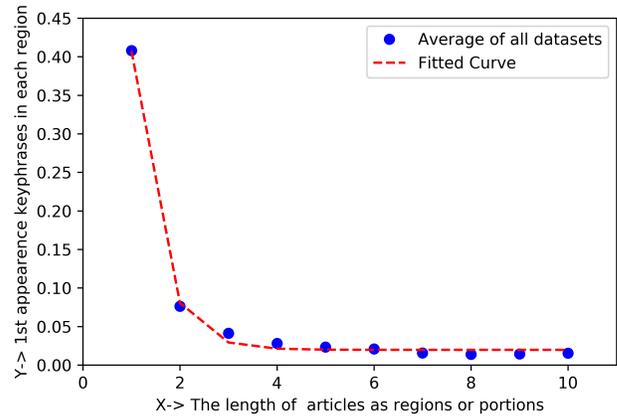


Fig. 7. The Curve Fitting Analysis of KCA Identification by Considering the Text Length as 10 Regions.

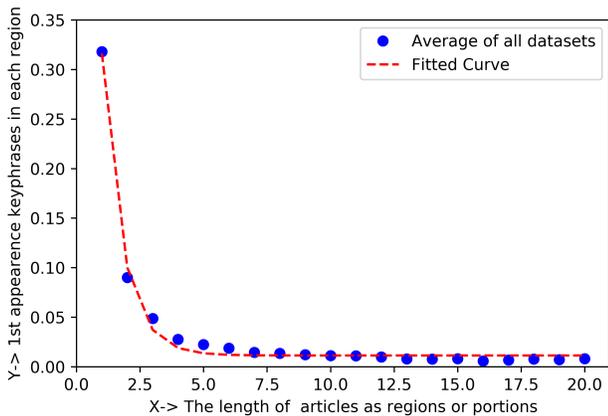


Fig. 6. The Curve Fitting Analysis of KCA Identification by Considering the Text Length as 20 Regions.

analysis of the curve fitting technique for KCA identification in each region is shown, with the text length divided into twenty (20) parts/regions, yielding the negative exponential equation expressed as follows (2) where  $p = 1.05$ ,  $q = 1.25$ , and  $r = 0.01$ . Similarly, KCA identification from this analysis for the length of text as ten (10) regions or portion is displayed in Fig. 7 and also gives the similar equation which is negative exponential in where  $p = 2.47$ ,  $q = 1.85$ , and  $r = 0.02$ . Since the fitted curves are found in negative exponential from the curve fitting analysis, It is demonstrated that most of the keyphrases are concentrated in the 1st portion of the documents, and next to the 2nd region of documents and so on, that are exhibited in Fig. 6 and Fig. 7.

$$y = p * e^{-qx} + r \quad (2)$$

### B. Comparison of Proposed Systems

Since KCA is a new technique with no existing policies, the proposed method does not compare with other techniques. The proposed system compares our two proposed approaches

considering the length of the documents as ten (10) regions and twenty (20) regions for KCA identification shown in the following Table III. Both proposed systems are employed 11 datasets for comparison. From Table III, in ten (10) regions, more keyphrases concentrated in 1st region (62.09%) than twenty (20) regions (48.37%) of the documents/articles. Similarly, in ten (10) regions, more keyphrases concentrated in 1st two regions combine (73.70%) than twenty (20) regions (62.08%) of the documents/articles. Afterwards, the ten(10) regions approach provides more keyphrases concentration in the 1st three regions combined (79.97%) than twenty (20) regions (69.48%). Finally, we can say that our proposed technique for ten (10) regions provide more keyphrase concentration than twenty (20) regions in 1st regions, then 2nd region, and so on. The KCA in an article is proven from these two approaches.

## VI. CONCLUSION

The extraction of features for the keyphrase extraction approach has evolved into a critical component in a wide range of computer science applications. A new unsupervised approach termed Keyphrases Concentrated Area identification as feature of keyphrase extraction is presented in this paper. It is domain and language independent, needs little statistical expertise, and does not need the use of train data. The proposed technique starts with data pre-processing, processing, and KCA identification (average calculation, plotting analysis, and curve-fitting analysis).The proposed approach effectively recognises the KCA from texts/articles and creates a negative exponential equation, showing that the first region of the document/article contains more keyphrases than the rest of the articles.

In comparison to the suggested two techniques, the system tested on 11 datasets and produced a superior result based on the 65.70 per cent existing goldkey. Taking use of the more statistical elements discussed in this research, we want to develop a strong keyphrase extraction approach in the future. Moreover, when multiple manually specified keywords are not found in the page, there are some limitations in resolving the missing goldkeys/keywords issue.

TABLE III. COMPARE OUR PROPOSED TWO APPROACHES FOR KCA IDENTIFICATION

| Articles Regions    | Keyphrase concentrated in 1st region(%) | Keyphrase concentrated in 1st two regions combine (%) | Keyphrase concentrated in 1st three regions combine (%) | Negative Exponential ( $p * e^{-q * x} + r$ ) |
|---------------------|-----------------------------------------|-------------------------------------------------------|---------------------------------------------------------|-----------------------------------------------|
| Ten (10) Regions    | 62.09%                                  | 73.70%                                                | 79.97%                                                  | p=2.47, q=1.85, r=0.02                        |
| Twenty (20) Regions | 48.37%                                  | 62.08%                                                | 69.48%                                                  | p=1.05, q=1.25, r=0.01                        |

#### ACKNOWLEDGMENT

The authors are grateful to University Malaysia Pahang for giving laboratory space and funding under the University FLAGSHIP Research Grants programme (Project number RDU192210 and RDU192212).

#### REFERENCES

- [1] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: a survey and trends," *Journal of Intelligent Information Systems*, vol. 54, no. 2, pp. 391–424, 2020.
- [2] C. Sun, L. Hu, S. Li, T. Li, H. Li, and L. Chi, "A review of unsupervised keyphrase extraction methods using within-collection resources," *Symmetry*, vol. 12, no. 11, p. 1864, 2020.
- [3] M. B. A. Miah, S. Awang, and M. S. Azad, "Region-based distance analysis of keyphrases: A new unsupervised method for extracting keyphrases feature from articles," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*. IEEE, 2021, pp. 124–129.
- [4] Z. A. Merrouni, B. Frikh, and B. Ouhbi, "Automatic keyphrase extraction: An overview of the state of the art," in *2016 4th IEEE international colloquium on information science and technology (CiSt)*. IEEE, 2016, pp. 306–313.
- [5] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel, "A semantic approach to contextual advertising," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 559–566.
- [6] F. Ricci, L. Rokach, and B. Shapira, "Introduction to recommender systems handbook," in *Recommender systems handbook*. Springer, 2011, pp. 1–35.
- [7] G. Rabby, S. Azad, M. Mahmud, K. Z. Zamli, and M. M. Rahman, "Teket: a tree-based unsupervised keyphrase extraction technique," *Cognitive Computation*, pp. 1–23, 2020.
- [8] Y. Sun, H. Qiu, Y. Zheng, Z. Wang, and C. Zhang, "Sifrank: a new baseline for unsupervised keyphrase extraction based on pre-trained language model," *IEEE Access*, vol. 8, pp. 10896–10906, 2020.
- [9] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 2, p. e1339, 2020.
- [10] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "Yake! keyword extraction from single documents using multiple local features," *Information Sciences*, vol. 509, pp. 257–289, 2020.
- [11] T. Li, L. Hu, H. Li, C. Sun, S. Li, and L. Chi, "Triplerank: An unsupervised keyphrase extraction algorithm," *Knowledge-Based Systems*, vol. 219, p. 106846, 2021.
- [12] T. Haarman, B. Zijlema, and M. Wiering, "Unsupervised keyphrase extraction for web pages," *Multimodal Technologies and Interaction*, vol. 3, no. 3, p. 58, 2019.
- [13] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "Yake! collection-independent automatic keyword extractor," in *European Conference on Information Retrieval*. Springer, 2018, pp. 806–810.
- [14] C. Florescu and C. Caragea, "Positionrank: An unsupervised approach to keyphrase extraction from scholarly documents," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1105–1115.
- [15] F. Boudin, "Unsupervised keyphrase extraction with multipartite graphs," *arXiv preprint arXiv:1803.08721*, 2018.
- [16] M. B. A. Miah and M. A. Yousuf, "Detection of lung cancer from ct image using image processing and neural network," in *2015 International conference on electrical engineering and information communication technology (ICEEICT)*. IEEE, 2015, pp. 1–6.
- [17] M. Al-Amin, M. B. Alam, and M. R. Mia, "Detection of cancerous and non-cancerous skin by using glcm matrix and neural network classifier," *International Journal of Computer Applications*, vol. 132, no. 8, p. 44, 2015.
- [18] M. B. A. Miah, "A real time road sign recognition using neural network," *International Journal of Computer Applications*, vol. 114, no. 13, 2015.
- [19] X. Meng, P. Zhang, Y. Xu, and H. Xie, "Construction of decision tree based on c4. 5 algorithm for online voltage stability assessment," *International Journal of Electrical Power & Energy Systems*, vol. 118, p. 105793, 2020.
- [20] E. Gopan, S. Rajesh, G. Vishnu, M. Thushara *et al.*, "Comparative study on different approaches in keyword extraction," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*. IEEE, 2020, pp. 70–74.
- [21] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *Proceedings of the 2009 conference on empirical methods in natural language processing*, 2009, pp. 1318–1327.
- [22] P. L. L. Romary, "Automatic key term extraction from scientific articles in grobid," in *SemEval 2010 Workshop*, 2010, p. 4.
- [23] M. Haddoud and S. Abdeddaïm, "Accurate keyphrase extraction by discriminating overlapping phrases," *Journal of Information Science*, vol. 40, no. 4, pp. 488–500, 2014.
- [24] F. Bulgarov and C. Caragea, "A comparison of supervised keyphrase extraction models," in *Proceedings of the 24th international conference on World Wide Web*, 2015, pp. 13–14.
- [25] F. Xie, X. Wu, and X. Zhu, "Efficient sequential pattern mining with wildcards for keyphrase extraction," *Knowledge-Based Systems*, vol. 115, pp. 27–39, 2017.
- [26] W. Souma, I. Vodenska, and L. Chitkushev, "Classification of paper values based on citation rank and pagerank," *Journal of Data and Information Science*, vol. 5, no. 3, p. 57, 2020.
- [27] C. Mallick, A. K. Das, M. Dutta, A. K. Das, and A. Sarkar, "Graph-based text summarization using modified textrank," in *Soft computing in data analytics*. Springer, 2019, pp. 137–146.
- [28] A. Bougouin, F. Boudin, and B. Daille, "Topicrank: Graph-based topic ranking for keyphrase extraction," in *International joint conference on natural language processing (IJCNLP)*, 2013, pp. 543–551.
- [29] X. Wan and J. Xiao, "Collabrank: towards a collaborative approach to single-document keyphrase extraction," in *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 2008, pp. 969–976.
- [30] L. Yao, Z. Pengzhou, and Z. Chi, "Research on news keyword extraction technology based on tf-idf and textrank," in *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE, 2019, pp. 452–455.
- [31] S. R. El-Beltagy and A. Rafea, "Kp-miner: Participation in semeval-2," in *Proceedings of the 5th international workshop on semantic evaluation*, 2010, pp. 190–193.
- [32] R. Campos and V. Mangaravite, "Datasets of automatic keyphrase extraction," 2020. [Online]. Available: <https://github.com/LIAAD/KeywordExtractor-Datasets>
- [33] O. Davydova, "Text preprocessing in python: Steps, tools, and examples," *Data Monsters* [https://es.wikipedia.org/wiki/Expresi%C3%B3n\\_regular](https://es.wikipedia.org/wiki/Expresi%C3%B3n_regular), 2019.

- [34] S. N. Kim, O. Medelyan, M.-Y. Kan, and T. Baldwin, "Semeval-2010 task 5: Automatic keyphrase extraction from scientific articles," in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 21–26.
- [35] L. Marujo, A. Gershman, J. Carbonell, R. Frederking, and J. P. Neto, "Supervised topical key phrase extraction of news stories using crowd-sourcing, light filtering and co-reference normalization," *arXiv preprint arXiv:1306.4886*, 2013.
- [36] T. D. Nguyen and M.-Y. Kan, "Keyphrase extraction in scientific publications," in *International conference on Asian digital libraries*. Springer, 2007, pp. 317–326.
- [37] A. T. Schutz *et al.*, "Keyphrase extraction from single documents in the open domain exploiting linguistic and statistical methods," *M. App. Sc Thesis*, 2008.
- [38] G. O. Aquino and L. C. Lanzarini, "Keyword identification in spanish documents using neural networks," *Journal of Computer Science & Technology*, vol. 15, 2015.
- [39] S. D. Gollapalli and C. Caragea, "Extracting keyphrases from research papers using citation networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.

# Transfer Learning based Performance Comparison of the Pre-Trained Deep Neural Networks

Jayapalan Senthil Kumar, Syahid Anuar, Noor Hafizah Hassan  
Razak Faculty of Technology and Informatics  
Universiti Teknologi Malaysia (UTM)  
54100 Kuala Lumpur, Malaysia

**Abstract**—Deep learning has grown tremendously in recent years, having a substantial impact on practically every discipline. Transfer learning allows us to transfer the knowledge of a model that has been formerly trained for a particular task to a new model that is attempting to solve a related but not identical problem. Specific layers of a pre-trained model must be retrained while the others must remain unmodified to adapt it to a new task effectively. There are typical issues in selecting the layers to be enabled for training and layers to be frozen, setting hyperparameter values, and all these concerns have a substantial effect on training capabilities as well as classification performance. The principal aim of this study is to compare the network performance of the selected pre-trained models based on transfer learning to help the selection of a suitable model for image classification. To accomplish the goal, we examined the performance of five pre-trained networks, such as SqueezeNet, GoogleNet, ShuffleNet, Darknet-53, and Inception-V3 with different Epochs, Learning Rates, and Mini-Batch Sizes to compare and evaluate the network's performance using confusion matrix. Based on the experimental findings, Inception-V3 has achieved the highest accuracy of 96.98%, as well as other evaluation metrics, including precision, sensitivity, specificity, and f1-score of 92.63%, 92.46%, 98.12%, and 92.49%, respectively.

**Keywords**—Transfer learning; deep neural networks; image classification; Convolutional Neural Network (CNN) models

## I. INTRODUCTION

The primary evolution of neural networks was stimulated by the desire to design a process that could imitate the human brain. The ability of conventional machine-learning approaches to explore natural data in its natural form was limited. Deep learning enables computational models with several processing layers to learn and represent data at multiple levels of abstraction, simulating how the brain receives and analyses multi-modal information, and so implicitly capturing intricate data structures [1]. A convolutional neural network (CNN) is one of the most popular deep learning models. It uses deep convolutional networks and non-linearity to discover local and spatial features, and patterns directly from raw data. As a result, a CNN learns features from data automatically, eliminating the necessity to manually extract them [2]. Image classification is a vital phenomenon in computer vision and other computer vision approaches, such as localisation, detection, and segmentation are built on top of it [3]. Deep neural networks (DNN) have recently been popular in the deep learning community for solving real-world issues, but the deep networks may face obstacles and hurdles throughout the training process, such as exploding/vanishing gradients and degradation [4]. The deep architecture presents the dedicated

concern of training a CNN from scratch, which needs massive computational power, a long training time, and a substantial amount of training data. The specificity of features rises as we progress from lower-level CNN layers to higher-level layers, until the last classification layer becomes profoundly task specific. The image features extracted by the lower-level CNN layers can be used to retrain the model for a completely different task, avoiding the need to start over [5]. In this case, all of the layers of a pre-trained CNN model can be employed as fixed feature extractors, with the exception of the final classification layer. Using the knowledge gained from earlier training, the final layer can be customised and retrained for a new task. When the depth of a network goes beyond the limit, it endures the degradation problem, which results in a decline in accuracy [6]. The internal covariate shift, which is the variation in the dissemination of the input data to a layer during training, is another matter of concern.

Transfer learning is a machine learning technique in which knowledge gained from one type of problem is applied to another similar task or domain [7]. CNN models are normally trained either from scratch or by applying transfer learning. Training from scratch involves a substantial amount of data to learn millions of parameters. Because a sufficiently labelled dataset is required for many applications, CNNs rarely train from scratch. Instead, a large-scale dataset is commonly used to pre-train a CNN, which is subsequently used as a fixed feature extractor or as an initialisation for other particular tasks [2]. The initial few layers of CNN models are trained to recognise task features. In the first layer, pre-trained models learn simple patterns like shapes and diagonals, then combine these components in successive layers to learn multipart features [8]. The models create meaningful constructs in the final layer by exploiting patterns learned from earlier layers. The final few layers of the trained network can be replaced and retrained with new layers for the target activity during transfer learning. Although fine tuned learning experimental studies need some learning, they are still much quicker than learning from the scratch [9], [6].

### A. Pre-Trained Deep Learning Architectures

The promotion of artificial neural networks (ANNs) is the deep neural network (DNN), which comprises numerous hidden layers between the input and output layers. A DNN is capable of expressing an object well through its deep architectures and excels at modelling complex nonlinear relationships [10]. In recent times, CNNs have played a critical role in image classification and object detection. In 1998

TABLE I. SUMMARY OF SELECTED PRE-TRAINED CNN MODELS

| Pre-Trained Models | Time | Depth | Layers | Image Input Size | Parameters |
|--------------------|------|-------|--------|------------------|------------|
| SqueezeNet [19]    | 2016 | 18    | 68     | 227-by-227       | 1.24 M     |
| GoogleNet [20]     | 2014 | 22    | 144    | 224-by-224       | 7.0 M      |
| ShuffleNet [21]    | 2018 | 50    | 173    | 224-by-224       | 1.4 M      |
| Darknet-53 [22]    | 2018 | 53    | 184    | 256-by-256       | 41.6 M     |
| Inception-V3 [23]  | 2016 | 48    | 315    | 299-by-299       | 23.9 M     |

LeCun et al. [11] proposed the first multilayer CNN, which is a convolutional network with seven levels that is simple to use, called LeNet-5. The layers of CNNs have become significantly deeper as GPU technology continues to advance. From 1998 to 2018, a number of CNN frameworks were developed, including LeNet [12], AlexNet [13], VGG 16 and VGG 19 [14], ResNet's Inception ResNet [15], ResNeXt, and other frameworks including PolyNet [16], DenseNet [17]. Transfer learning at a deep level instead of utilising traditional machine learning approaches that benefit from handcrafted features, CNN learns the most representative features from raw data automatically [18]. A variety of CNN architecture modifications with a rapid growth in the number of layers have recently been demonstrated. In this study, five pre-trained models, namely SqueezeNet, GoogleNet, ShuffleNet, Darknet-53, and Inception-v3 have been selected for the performance comparison and a brief summary is provided in Table I.

A system designer must incorporate their judgment and substantial feature engineering to resolve the question of what needs to be transferred. The challenge on how knowledge should be conveyed through is model selection and how to supplement it to enhance prediction performance [11]. When selecting a network to apply to a problem, different aspects of pre-trained models are important to consider. Network accuracy, speed, and size are the most important considerations. Choosing a network is usually a compromise between these factors. The primary goal of this study is to compare the network performance of the selected pre-trained models based on accuracy, speed, and size to help the selection of a suitable model for image classification.

The rest of the paper is organised as follows. In Section II, we give a description of the related works. In Section III, the methodology is described in detail together with the transfer learning steps used in MATLAB. In Section IV, the experimental results are shown, followed by the performance evaluation and performance comparison of the pre-trained deep neural networks. Finally, in Section V, we provide the conclusions and future work.

## II. RELATED WORK

In several disciplines, traditional machine learning algorithms have been widely accepted. Deep learning as well as image processing techniques have been used. With the introduction of transfer learning as a new learning framework [24], by fine-tuning pre-trained CNN models that have already been trained on ImageNet, similar results can now be obtained on deep learning applications. These models require a smaller number of training examples than developed models, which necessitate a significant amount of effort to acquire a big number

of training instances [25]. Transfer learning has been used in a variety of fields, including agriculture, where it has been used to identify weeds, classify land cover, identify plants, count fruits, and classify crop types. Transfer learning has become increasingly important in medical image processing, while pre-trained deep neural networks have made significant advances in the medical field, including the use of magnetic resonance imaging (MRI) scans, computerised tomography (CT) scans, and electrocardiograms (ECs) to detect life-threatening diseases, such as heart disease, cancer, and brain tumours. Shakil Ahmed et. al. [8] developed a transfer learning-based framework, which was tested against two well-known CNN models, Inception-V3 and VGG-16, using the Kimia Path24 dataset, which was created specifically for the classification and retrieval of histopathological images. Muhammed Talo [26] did the same kind of study with the same Kimia Path24 dataset. ResNet-50 and DenseNet-161, however, were used as well-known pre-trained CNN models. Rishav Singh et. al [11] presented a framework based on the concept of transfer learning to address and focus efforts on histopathology and unbalanced image classification, employing the widely used VGG-19 as a base model.

Samuel Kumaresan et. al. [18], suggested employing transfer learning to overcome the issue of a small dataset of welding defect X-ray pictures. They used two large pre-trained convolutional neural networks, VGG16 and ResNet50, to extract features from weld defect radiograph images that can be used to classify 14 different types of weld defects. The goal for Edna Chebet Too et. al. [6] was to fine-tune and explore the deep convolutional neural network for image-based plant disease classification. The models VGG 16, Inception V4, ResNet with 50, 101, and 152 layers, and DenseNet with 121 layers were assessed in an empirical comparison of the deep learning architectures. Jianping Ju et. al. [27] in an effort to address the actual demand for jujube fault detection, introduced a jujube sorting model in small data sets based on convolutional neural networks and transfer learning using the SE-ResNet50-TL and SE-ResNet50-CL models. Triplet loss function and Center loss function were used to replace SoftMax loss function and embedded SE module for the dry red date defect detection. Alper et. al. [28] recommended a new CNN architecture for the hazelnut variety classification and the model was compared with four pre-trained models: VGG16, VGG19, InceptionV3, and ResNet50. In recent years, the difficulty of layer selection when using transfer learning with fine-tuning has received substantial attention. With the widespread adoption of deep learning techniques, transfer learning with fine-tuning appeared to be the ultimate approach for transferring knowledge, allowing scientists and professionals to apply such deep learning methods more quickly to a variety of domain problems [29].

## III. METHODOLOGY

Classification has been an effective mission, and it is important in the subject of computer vision, which seeks to classify images into predefined classes automatically. Prior to the boom of deep learning approaches, a lot of effort was invested into constructing scale-invariant features, feature representations, and image classification classifiers [30]. These well-crafted qualities, on the other hand, work against objects in natural images with complex scenes, varying colour, texture, and illumination, as well as constantly changing positions

and view parameters. Researchers have been working on sophisticated ways to increase image classification accuracy for decades. When the large-scale image dataset ImageNet was formed in 2009, Feifei Li [31] created the great-leap-forward advancement of image classification. The information about dataset, learning environment, gradient, learning rate, epoch, and mini-batch size employed in MATLAB, and the steps of transfer learning used in this study are explained under the methodology in the subsequent sections.

#### A. Dataset

The CIFAR-10 [32] dataset has 32 x 32 colour images that are divided into ten classes, each with 5,000 training images and 1,000 test images. Among the ten classes, five classes have been selected for the experimental process. From the training dataset, 3,000 training images are selected for each of the five classes, which includes the list as shown below:

Selected Classes = ('bird', 'cat', 'deer', 'dog', 'horse');

The most widely used split ratios are 70:30; 80:20; 65:35; 60:40 etc., in which the sample size suits the nature of the problem and the architecture implemented. There is no fixed law for dividing training and test datasets when it comes to data splitting. Some scholars have traditionally used the 70:30 ratio to differentiate the datasets. As most widely used in MATLAB, the training set of images is split into training set and validation set by the 70:30 ratio.

```
[imdsTrain, imdsValidation] = splitEachLabel(imds, 0.7);
```

Besides that, image augmentation could be used at random on the training datasets with distinct values to help expand the dataset, preventing the network from overfitting and capturing the exact features of the training images. The following settings for image augmentation have been chosen, as indicated in most of the MATLAB examples: horizontal reflection, horizontal and vertical translation in the range [-30 30] pixels, and horizontal and vertical scaling in the range [0.9 1.1] with a random rate.

#### B. Learning of the Pre-Trained Networks

- Environment – The networks are implemented in MATLAB R2021a. The size of input images is adjusted to match the layers of various models.
- Stochastic Gradient Descent with Momentum (SGDM) – Gradient descent [10] is a popular neural network optimisation approach that can tackle a variety of trivial issues. When the training dataset is huge, however, the simple gradient descent method may use a lot of processing resources, making the convergence process slow. Simultaneously, because the gradient descent approach considers all of the training data for each calculation, it may result in overfitting. To resolve this challenging dispute, SGDM has been considered in this study. Momentum [33] is a commonly used acceleration technique in the gradient descent method whereby the convergence process can be accelerated.
- Learning Rate (LR) – When it comes to CNN training, LR is a crucial parameter. The LR is frequently

decreased by a factor of 0.1 or 0.5. In this study, fixed learning rates of LR-0.001 and LR-0.0001 have been chosen instead of reducing the LR by each epoch.

- Epoch – The complete pass of the training algorithm across the entire training set is referred to as an epoch. In this study, the selected epoch values are 10, 20, 30.
- Mini Batch Size – A mini-batch is a subset of the training set that is utilised to calculate the loss function's gradient and update the weights. Two batch sizes are selected as part of the experimentation process: 32 and 64.

#### C. Transfer Learning Flow in MATLAB

CNN's unique qualities, such as incremental feature extraction in subsequent layers, make it possible to use parts of a pre-trained model for a completely new task without retraining the entire network [34], [35]. The fundamental idea is to use the initial layers from a pre-trained model and just retrain the last few layers on new images. Transfer learning can be implemented by replacing the output layer with a new classifier and then, selecting one of the two approaches:

- Fixed feature extraction by freezing the initial layers or other layers of the convolutional base.
- Fine-tuning the weights and other parameters to retrain one or more convolution layers.

The entire experimentation process of image classification with the dataset CIFAR-10 has been done with MATLAB. The CIFAR-10 dataset is downloaded and provided as input data to the pre-trained model. Prior to loading the data, the entire dataset is divided into three main datasets comprising the training, validation, and testing datasets. To get good performance, deep neural networks require a vast amount of training data. Image augmentation, such as reflection, translation, and scaling, are used to increase the performance of deep networks in order to develop an effective image classifier with little training data. The pre-trained network is loaded, and the final layers are replaced with a new classification layer and a fully connected or convolutional layer. To fine-tune the model, the initial layers of each network are frozen and other parameters like the pool size, stride etc., are updated. The freezing layers are chosen according to the depth, size, and number of layers of the pre-trained network. Afterwards, the training process is initiated, followed by the classification of the validation, and test images. Finally, the classification accuracy is computed and performance of the networks are evaluated using confusion matrix. The transfer learning steps are illustrated in Fig. 1 which is then followed by the detailed steps performed in MATLAB to implement the entire transfer learning process.

First, training and validation sets are used to perform the training process to make sure that we have the best possible training model in the study. By checking the accuracy and loss of training and validation sets, we will be able to control the model's performance during training. Thus, the best possible model can be obtained by fine-tuning at the end of the training process. The results are evaluated by using the test set and these procedures were applied for each of the five models used in the study. All parameters are used in the same way for each model and the models used were evaluated using confusion

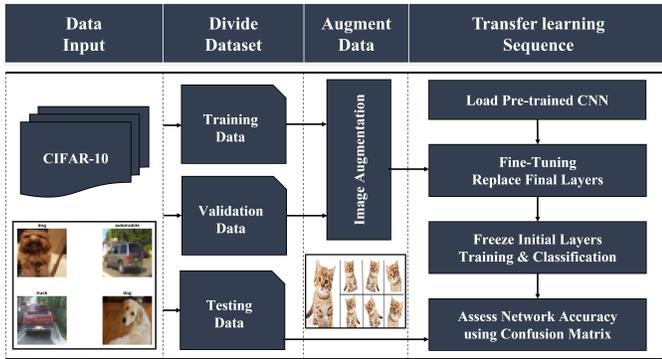


Fig. 1. Flow of Transfer Learning Sequence.

**Transfer Learning Steps - MATLAB**

**Prepare Data**

- 1) Downloading the data.
- 2) <https://www.cs.toronto.edu/~kriz/cifar.html>.

**Load Data**

- 3) Selection of classes from the downloaded data.
- 4) `imds = imageDatastore(fullfile(rootFolder));`

**Load Pretrained Network**

- 5) `net = SqueezeNet | GoogleNet | ShuffleNet | Darknet-53 | Inception-v3;`
- 6) Analyze the network.

**Replace Final Layers**

- 7) `lgraph = layerGraph(net);`
- 8) `lgraph = replaceLayer(lgraph, learnableLayer.Name, newLearnableLayer);`

**Freeze Initial Layers**

- 9) `layers = lgraph.Layers;`
- 10) `layers(1:10) = freezeWeights(layers(1:10));`

**Train Network**

- 11) Training options: `['MiniBatchSize', 'MaxEpochs', 'InitialLearnRate'];`
- 12) `net = trainNetwork(augimdsTrain, lgraph, options);`

**Classify Images**

- 13) Validation, Testing.

**Accuracy and Loss Plot**

- 14) Validation.

**Confusion Matrix**

- 15) `cm = confusionmat(trueLabels, predictedLabels);`
- 16) `cm_chart = confusionchart(trueLabels, predictedLabels);`

**Reset GPU**

matrix to find out the performance of the classifier. The performance evaluation used in the study and the comparison on the performance of different models are presented in Table VIII and Table IX in the following section.

**IV. RESULT AND DISCUSSION**

The whole experimental process was carried out with a laptop and the experimental setup including the hardware, software, and its specifications are mentioned in Table II.

**A. Experimental Results**

In the training process, each iteration involves a gradient estimation and a network parameter update. Training can be tracked in MATLAB to determine how quickly the network's accuracy improves, as well as if the network attempts to overfit the training data. Following the completion of the training, the results can be inspected to see the finalised validation accuracy and to discover how the training was proceeded by plotting the key metrics, which include training accuracy, validation

TABLE II. EXPERIMENTAL SETUP

| Hardware/Software       | Specifications                              |
|-------------------------|---------------------------------------------|
| Microprocessor          | AMD Ryzen 7 5800H- Radeon Graphics@3.20 GHz |
| RAM                     | 16.0 GB                                     |
| GPU                     | NVIDIA GeForce RTX 3060 Laptop GPU          |
| Dedicated Video RAM     | 6.0 GB                                      |
| Deep Learning Framework | MATLAB R2021a – 64 bit                      |
| Programming Language    | MATLAB                                      |
| Operating System        | Windows 10 Home Single Language             |

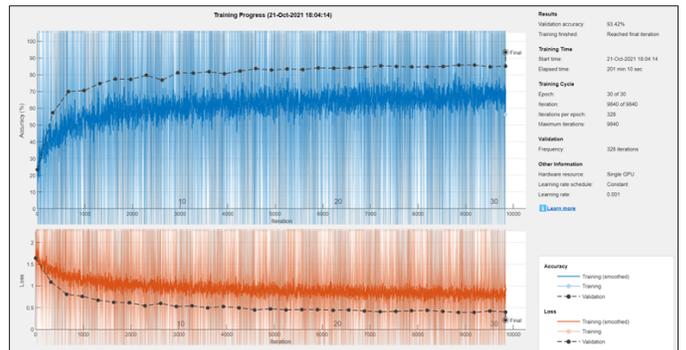


Fig. 2. Training Progress Sample – MATLAB.

accuracy, training loss, and validation loss. Fig. 2 depicts the sample of training progress accomplished with MATLAB, which primarily highlights the results for validation accuracy, training time, training cycle with iterations and epoch, validation, and about the hardware resources. The validation plots are portrayed in Fig. 3, and they contain the validation accuracy, which represents the classification accuracy, and the validation loss, which represents the validation loss across the entire validation set for the five pre-trained networks.

The experimental results are presented in the tables

TABLE III. SQUEEZENET – FREEZING LAYERS:[1-11]

| Hyper Parameters              | Validation Accuracy (%) | Test Accuracy (%) | Time (mins) |
|-------------------------------|-------------------------|-------------------|-------------|
| <b>LR - 0.001 Epoch - 30</b>  |                         |                   |             |
| Mini Batch Size- 64           | 81.87                   | 79.10             | 21.20       |
| Mini Batch Size- 32           | 81.40                   | 78.90             | 20.15       |
| <b>LR - 0.0001 Epoch - 30</b> |                         |                   |             |
| Mini Batch Size- 64           | 72.73                   | 70.78             | 21.19       |
| Mini Batch Size- 32           | 78.36                   | 76.28             | 20.17       |

TABLE IV. GOOGLNET – FREEZING LAYERS:[1-10]

| Hyper Parameters              | Validation Accuracy (%) | Test Accuracy (%) | Time (mins) |
|-------------------------------|-------------------------|-------------------|-------------|
| <b>LR - 0.001 Epoch - 30</b>  |                         |                   |             |
| Mini Batch Size- 64           | 89.29                   | 88.58             | 57.28       |
| Mini Batch Size- 32           | 90.16                   | 88.82             | 53.37       |
| <b>LR - 0.0001 Epoch - 30</b> |                         |                   |             |
| Mini Batch Size- 64           | 83                      | 83.94             | 58.21       |
| Mini Batch Size- 32           | 85.98                   | 86.10             | 53.51       |

TABLE V. SHUFFLENET – FREEZING LAYERS:[1-15]

| Hyper Parameters              | Validation Accuracy (%) | Test Accuracy (%) | Time (mins) |
|-------------------------------|-------------------------|-------------------|-------------|
| <b>LR - 0.001 Epoch - 30</b>  |                         |                   |             |
| Mini Batch Size- 64           | 88.93                   | 87.18             | 97.21       |
| Mini Batch Size- 32           | 88.07                   | 85.08             | 111.29      |
| <b>LR - 0.0001 Epoch - 30</b> |                         |                   |             |
| Mini Batch Size- 64           | 79.33                   | 78.26             | 98.70       |
| Mini Batch Size- 32           | 85.02                   | 82.54             | 113.11      |

TABLE VI. DARKNET-53 – FREEZING LAYERS:[1-14]

| Hyper Parameters              | Validation Accuracy (%)     | Test Accuracy (%) | Time (mins) |
|-------------------------------|-----------------------------|-------------------|-------------|
| <b>LR - 0.001 Epoch - 30</b>  |                             |                   |             |
| Mini Batch Size- 64           | Error@19/30 (Out of Memory) |                   |             |
| Mini Batch Size- 32           | 89.89                       | 86.62             | 207.13      |
| <b>LR - 0.0001 Epoch - 30</b> |                             |                   |             |
| Mini Batch Size- 64           | Error@19/30 (Out of Memory) |                   |             |
| Mini Batch Size- 32           | 90.58                       | 86.76             | 203.30      |

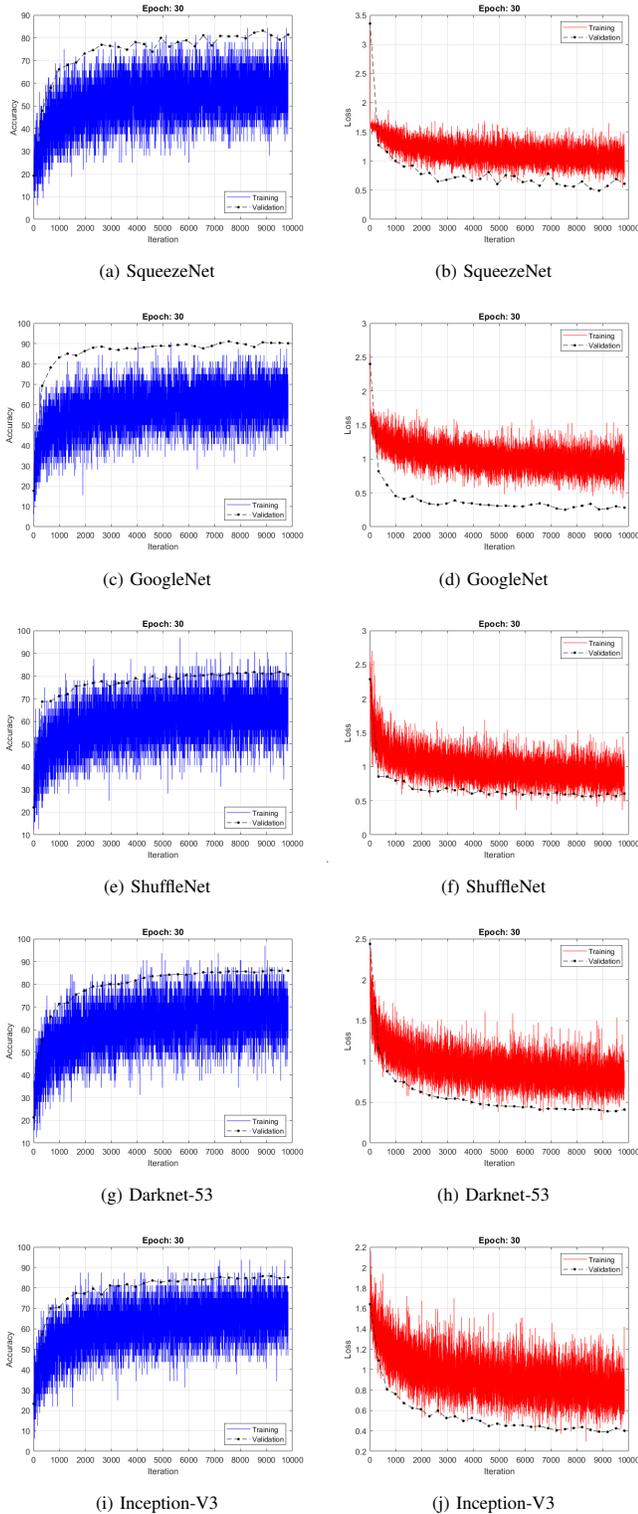


Fig. 3. Validation Plots - Batch-32.

such as Table III-SqueezeNet, Table IV-GoogleNet, Table V-ShuffleNet, Table VI-Darknet-53 and Table VII-Inception-V3, including the hyperparameters such as mini-batch size, learning rate (LR), epoch as well as the validation accuracy and testing accuracy with the elapsed time to complete the training progress. The experimental findings made it possible to emphasise the following outcomes,

- Epoch-30 was chosen for further comparison based on the experimental findings, and the results were quite promising.
- When it comes to mini batch sizes, batch 32 has shown to be more promising than batch 64. Also, with Darknet-53, batch 64 displayed an error due to a lack of RAM (out of memory); hence batch 32 was chosen for further evaluation and comparison.
- With the exception of Darknet-53, LR-0.001 yielded favourable results when compared to LR-0.0001. For the subsequent studies, LR-0.001 findings were chosen for the other four networks, and LR-0.0001 results for Darknet-53.
- Out of the five pre-trained networks, Inception-V3 produced the best results, with the most layers, while SqueezeNet produced the unpleasant results, with the

TABLE VII. INCEPTION-V3 – FREEZING LAYERS:[1-41]

| Hyper Parameters              | Validation Accuracy (%) | Test Accuracy (%) | Time (mins) |
|-------------------------------|-------------------------|-------------------|-------------|
| <b>LR - 0.001 Epoch - 30</b>  |                         |                   |             |
| Mini Batch Size- 64           | 92.78                   | 91.60             | 165.16      |
| Mini Batch Size- 32           | 93.42                   | 92.46             | 201.10      |
| <b>LR - 0.0001 Epoch - 30</b> |                         |                   |             |
| Mini Batch Size- 64           | 80.29                   | 76.54             | 165.80      |
| Mini Batch Size- 32           | 87.53                   | 83.86             | 206.70      |

fewest layers.

**B. Performance Evaluation using Confusion Matrix**

The ratio between the number of right predictions made and the total number of predictions produced is known as classification accuracy [18]. The learning performance of the pre-trained deep neural networks is assessed using a standard confusion matrix method. A confusion matrix is a summary of classification problem prediction outcomes. It provides insight into correct and incorrect classifications, as well as the types of errors made, for each specific class. In image classification, the confusion matrix is primarily used to compare the classification to the actual measurement value in order to intuitively and accurately describe the accuracy of model classification [36]. The confusion matrix can be used to directly identify the performance of deep CNN models, and the evaluation metrics are listed below:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where ACC stands for accuracy, which is defined as the percentage of correctly classified samples when a measured value is compared to a known value.

$$PREC = \frac{TP}{TP + FP} \quad (2)$$

where PREC is the precision used to determine the model’s ability to correctly classify positive values.

$$SENS = \frac{TP}{TP + FN} \quad (3)$$

where SENS is the sensitivity, also known as recall, which is the frequency with which the model correctly predicts positive values. It’s used to figure out how well the model can predict positive values.

$$SPEC = \frac{TP}{TN + FP} \quad (4)$$

where SPEC denotes the specificity with which the model’s ability to predict negative values.

$$F1 - Score = \frac{2 * PREC * SENS}{PREC + SENS} \quad (5)$$

whereas the harmonic mean of the precision and sensitivity is the F1-score, also known as the balanced F-score or F-measure.

In MATLAB, the predicted class is represented by the rows, while the true class is represented by the columns. The diagonal cells relate to accurately classified observations. The off-diagonal cells correspond to observations that were inaccurately classified. The number of accurately and inaccurately classified observations for each predicted class are displayed as percentages of the total number of observations in the respective predicted class in a column-normalized column summary. The number of accurately and inaccurately classified observations for each true class are displayed as percentages of the total number of observations for that true class in a row-normalized row summary.

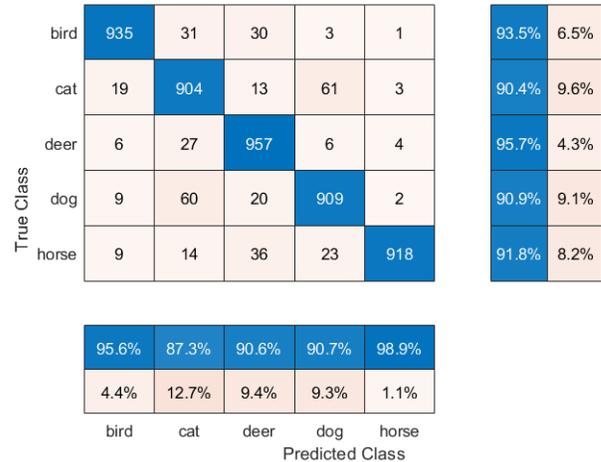


Fig. 4. Confusion Matrix of Inception-v3 - LR-0.001.

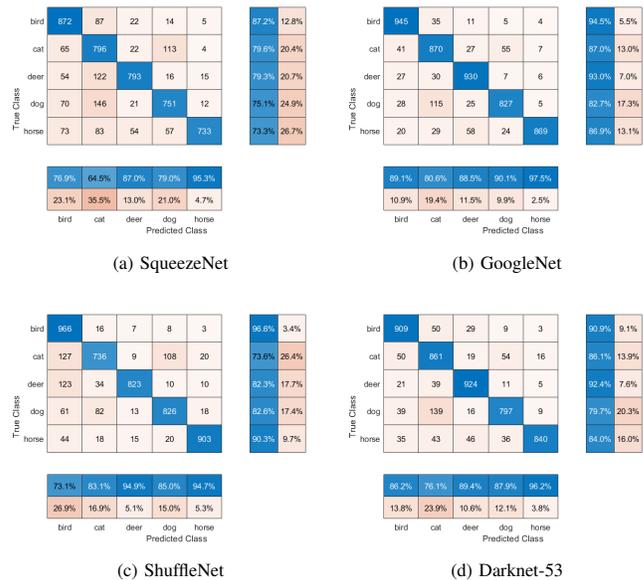


Fig. 5. Confusion Matrices for LR-0.001|Batch-32.

The Fig. 4 demonstrates the confusion matrix for the pre-trained network Inception-V3 and the Fig. 5 represents the confusion matrices for networks such as (a) SqueezeNet, (b) GoogleNet, (c) ShuffleNet and (d) Darknet-53 for the mini-batch 32 and LR-0.001. The Fig. 6 describes the confusion matrix for the pre-trained network Darknet-53 and the Fig. 7 represents the confusion matrices for networks such as (a) SqueezeNet, (b) GoogleNet, (c) ShuffleNet and (d) Inception-V3 for the mini-batch 32 and LR-0.0001. When it came to learning rates, among the selected five pre-trained networks LR-0.0001 was outperformed by LR-0.001. Under LR-0.001, almost all of the networks performed well in classifying the images and prediction, however under LR-0.0001, most of the networks struggled to predict the positive values. Based on the confusion matrices, the following inferences were discovered,

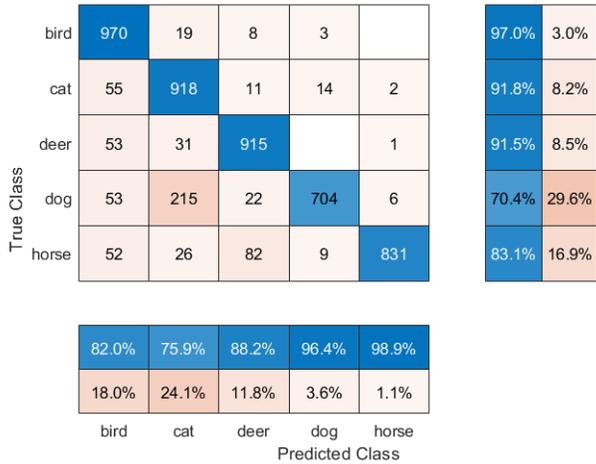


Fig. 6. Confusion Matrix of Darknet-53 - LR-0.0001.

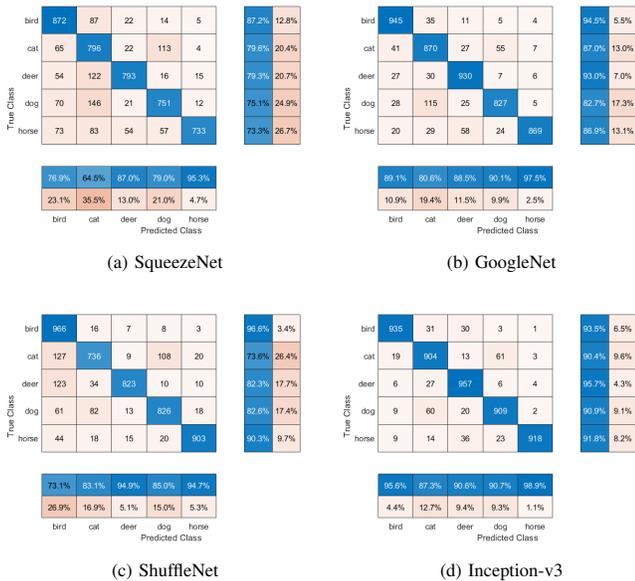


Fig. 7. Confusion Matrices for LR-0.0001|Batch-32.

- In terms of LR – 0.001, Inception-V3 outperformed the rest of the pre-trained networks in the model’s ability to predict positive values followed by GoogleNet and Darknet-53. With Inception-V3, all five classes were correctly classified with an overall accuracy of above 90%. Among the classes deer class made the highest score whereas 957 out of 1000 images were classified correctly. When it comes to prediction, the horse class scored high of correctly predicting 98.9% of the positive values. For all the five classes, GoogleNet scored 80% or higher, with the horse class getting the highest prediction score, predicting 97.5% of positive values. Darknet-53 scored the highest among the networks a prediction score of 96.2% in the horse class but got a very least score of only 76.1% in the cat class. ShuffleNet had a mediocre

TABLE VIII. EVALUATION RESULTS OF THE PRE-TRAINED NETWORKS

| Pre-Trained Models            | ACC (%)      | PREC (%)     | SENS (%)     | SPEC (%)     | F1-Score (%) |
|-------------------------------|--------------|--------------|--------------|--------------|--------------|
| <b>LR - 0.001 Epoch - 30</b>  |              |              |              |              |              |
| SqueezeNet                    | 91.56        | 80.53        | 78.90        | 94.73        | 79.16        |
| GoogleNet                     | 95.53        | 89.16        | 88.82        | 97.21        | 88.85        |
| ShuffleNet                    | 94.03        | 86.15        | 85.08        | 96.27        | 85.13        |
| Darknet-53                    | 94.65        | 87.15        | 86.62        | 96.65        | 86.68        |
| <b>Inception-V3</b>           | <b>96.98</b> | <b>92.63</b> | <b>92.46</b> | <b>98.12</b> | <b>92.49</b> |
| <b>LR - 0.0001 Epoch - 30</b> |              |              |              |              |              |
| SqueezeNet                    | 90.51        | 77.89        | 76.28        | 94.07        | 76.14        |
| GoogleNet                     | 94.44        | 86.14        | 86.10        | 96.53        | 86.06        |
| ShuffleNet                    | 93.02        | 82.76        | 82.54        | 95.63        | 82.57        |
| <b>Darknet-53</b>             | <b>94.70</b> | <b>88.29</b> | <b>86.76</b> | <b>96.69</b> | <b>86.70</b> |
| Inception-V3                  | 93.54        | 84.55        | 83.86        | 95.96        | 83.91        |

performance, scoring 94.7% in the horse class and a very low prediction score of 73.1% in the bird class. SqueezeNet had the lowest classification performance of all the networks, with a prediction score of 95.3% in the horse category and 64.5% in the cat category.

- In terms of LR – 0.0001, Darknet-53 outperformed the rest of the pre-trained networks in the model’s ability to predict positive values followed by Inception-V3 and GoogleNet. If compared with LR 0.001, Darknet-53 scored the maximum classification accuracy under LR 0.0001 with the highest score of 97% among all the networks. With prediction, scored the highest of 98.9% in horse class and 75.9% in cat class. The bird class received the highest classification score of 93% in Inception-V3, with 930 out of 1000 images correctly classified, whereas the model struggled to predict the positive values of the bird class, accounting for 77.1%. In GoogleNet among the five classes, horse class got the highest prediction score of 91.1% and cat class got the lowest score of 81.7%. ShuffleNet had an average classification compared to other networks whereas horse class got the highest prediction score of 91.5% and cat class got the lowest score of 76.1%. SqueezeNet had the lowest performance among all other networks with the lowest prediction score of 66.0% for the bird class whereas got the highest score of 90.7% for the horse class.

The results of the classification metrics evaluation of the five pre-trained networks for both the Learning Rates of 0.001 and 0.0001 are summarized in Table VIII. According to the evaluation results it is evident that the networks performed well on the LR-0.001 in compared to LR-0.0001 except for Darknet-53. Darknet-53, in compared to the other four networks, showed promising results with a LR-0.0001, whilst the other networks’ performances were on the decline.

### C. Pre-Trained Networks Performance Comparison

The performance comparison of the five pre-trained networks, encompassing both LR, is shown in Table IX. Based on the performance comparison of the pre-trained networks, Inception-V3 has achieved the highest accuracy of 96.98%, as well as other metrics such as precision, sensitivity, specificity,

TABLE IX. PERFORMANCE COMPARISON OF THE PRE-TRAINED NETWORKS

| Pre-Trained Models  | ACC (%)      | PREC (%)     | SENS (%)     | SPEC (%)     | F1-Score (%) |
|---------------------|--------------|--------------|--------------|--------------|--------------|
| SqueezeNet          | 91.56        | 80.53        | 78.90        | 94.73        | 79.16        |
| GoogleNet           | 95.53        | 89.16        | 88.82        | 97.21        | 88.85        |
| ShuffleNet          | 94.03        | 86.15        | 85.08        | 96.27        | 85.13        |
| Darknet-53          | 94.70        | 88.29        | 86.76        | 96.69        | 86.70        |
| <b>Inception-V3</b> | <b>96.98</b> | <b>92.63</b> | <b>92.46</b> | <b>98.12</b> | <b>92.49</b> |

and f1-score. The other networks produced somewhat lower results than Inception-V3, but altogether, all five pre-trained networks attained an accuracy of 90% or higher.

Transfer learning using a pre-trained CNN model is a better option for classification with the availability of only small datasets. In many of the previous studies, different pre-trained CNN models were compared using medical images and other relevant datasets. The results showed that the performance of the pre-trained models were mainly based on the dataset. With the limited computing resources, only five classes of the benchmark CIFAR-10 dataset are selected, but still managed to accomplish the aim of this study in the selection of a suitable model for image classification, Inception-V3. Even though one of the CNN model darknet-53 produced an error (out of memory) over batch-64, the current study shows that pre-trained networks with the highest number of layers (Darknet-53 and Inception-V3) provided the maximum scores in the prediction of classes with best accuracy. The current study proves that transfer learning can be useful for various computer vision problems, especially for the ones with small datasets. With the availability of proper datasets, deep CNN models have the capabilities to take medical imaging technology further, providing a higher level of automation in medical imaging, including image processing and analysis.

## V. CONCLUSIONS

In this study, we experimented with the performance of five pre-trained networks, such as SqueezeNet, GoogleNet, ShuffleNet, Darknet-53, and Inception-V3 with different epochs, learning rates, and mini-batch sizes. We performed the entire training process in MATLAB R2021a where we can view the complete network architecture of the CNN models, which helped us in the selection of freezing the initial layers. The final layers of the pre-trained CNN models are replaced either with a fully connected layer or convolutional layer and a new classifier replacing the classification layer. The initial layers are frozen to keep the weights intact and after the training, each model was evaluated using a confusion matrix. The experimental findings show that each pre-trained network produced different results with different hyper-parameters in the prediction of positive values. The results demonstrate that all the five pre-trained networks yielded promising results over the mini batch size-32, and epoch-30. In terms of LR, Darknet-53 delivered impressive results with LR-0.0001, achieving a maximum accuracy of 94.70%. Overall, the Inception-V3 model with LR-0.001 achieved the highest accuracy of 96.98%, as well as other evaluation metrics including precision, sensitivity, specificity, and f1-score of 92.63%, 92.46%, 98.12%, and 92.49%, respectively.

## VI. FUTURE WORK

We presented a transfer learning-based performance comparison between the selected five pre-trained networks in this study. The freezing of network layers was selected based on the network depth, size, and number of layers. Only the initial layers were frozen; however, different sets of layers can be frozen. In the future, focus will be given to freeze multiple set of layers and to compare the results of frozen and non-frozen layers of the pre-trained networks. For further evaluation and comparison, different datasets and other pre-trained deep neural networks can also be explored.

## ACKNOWLEDGMENT

The authors are grateful to Universiti Teknologi Malaysia (UTM) for supporting the work with cost center number Q.K130000.2656.17J22.

## REFERENCES

- [1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.
- [2] H. Altaheri, M. Alsulaiman, and G. Muhammad, "Date fruit classification for robotic harvesting in a natural environment using deep learning," *IEEE Access*, vol. 7, pp. 117 115–117 133, 2019.
- [3] W. Rawat and Z. Wang, "Deep convolutional neural networks for image classification: A comprehensive review," *Neural computation*, vol. 29, no. 9, pp. 2352–2449, 2017.
- [4] K. Goutam, S. Balasubramanian, D. Gera, and R. R. Sarma, "Layerout: Freezing layers in deep neural networks," *SN Computer Science*, vol. 1, no. 5, pp. 1–9, 2020.
- [5] U. A. Khan, M. A. Martinez-Del-Amor, S. M. Altowaijri, A. Ahmed, A. U. Rahman, N. U. Sama, K. Haseeb, and N. Islam, "Movie tags prediction and segmentation using deep learning," *IEEE Access*, vol. 8, pp. 6071–6086, 2020.
- [6] E. C. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture*, vol. 161, pp. 272–279, 2019.
- [7] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [8] S. Ahmed, A. Shaikh, H. Alshahrani, A. Alghamdi, M. Alrizq, J. Baber, and M. Bakhtyar, "Transfer learning approach for classification of histopathology whole slide images," *Sensors*, vol. 21, no. 16, p. 5361, 2021.
- [9] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in plant science*, vol. 7, p. 1419, 2016.
- [10] C. Lu and W. Li, "Ship classification in high-resolution sar images via transfer learning with small training dataset," *Sensors*, vol. 19, no. 1, p. 63, 2019.
- [11] R. Singh, T. Ahmed, A. Kumar, A. K. Singh, A. K. Pandey, and S. K. Singh, "Imbalanced breast cancer classification using transfer learning," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 18, no. 1, pp. 83–93, 2020.
- [12] Y. LeCun *et al.*, "Lenet-5, convolutional neural networks," *URL: http://yann.lecun.com/exdb/lenet*, vol. 20, no. 5, p. 14, 2015.
- [13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [15] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

- [16] X. Zhang, Z. Li, C. Change Loy, and D. Lin, "Polynet: A pursuit of structural diversity in very deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 718–726.
- [17] G. Huang, S. Liu, L. Van der Maaten, and K. Q. Weinberger, "Condensenet: An efficient densenet using learned group convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2752–2761.
- [18] S. Kumaresan, K. J. Aultrin, S. Kumar, and M. D. Anand, "Transfer learning with cnn for classification of weld defect," *IEEE Access*, vol. 9, pp. 95 097–95 108, 2021.
- [19] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and 0.5 mb model size," *arXiv preprint arXiv:1602.07360*, 2016.
- [20] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [21] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.
- [22] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [23] 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.
- [24] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International conference on artificial neural networks*. Springer, 2018, pp. 270–279.
- [25] M. Boulares, T. Alafif, and A. Barnawi, "Transfer learning benchmark for cardiovascular disease recognition," *IEEE Access*, vol. 8, pp. 109 475–109 491, 2020.
- [26] M. Talo, "Automated classification of histopathology images using transfer learning," *Artificial Intelligence in Medicine*, vol. 101, p. 101743, 2019.
- [27] J. Ju, H. Zheng, X. Xu, Z. Guo, Z. Zheng, and M. Lin, "Classification of jujube defects in small data sets based on transfer learning," *Neural Computing and Applications*, pp. 1–14, 2021.
- [28] A. Taner, Y. B. Öztekin, and H. Duran, "Performance analysis of deep learning cnn models for variety classification in hazelnut," *Sustainability*, vol. 13, no. 12, p. 6527, 2021.
- [29] G. Vrbančič and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *IEEE Access*, vol. 8, pp. 196 197–196 211, 2020.
- [30] X. Feng, Y. Jiang, X. Yang, M. Du, and X. Li, "Computer vision algorithms and hardware implementations: A survey," *Integration*, vol. 69, pp. 309–320, 2019.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
- [32] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [33] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.
- [34] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *arXiv preprint arXiv:1411.1792*, 2014.
- [35] A. Brodzicki, J. Jaworek-Korjakowska, P. Kleczek, M. Garland, and M. Bogyo, "Pre-trained deep convolutional neural network for clostridoides difficile bacteria cytotoxicity classification based on fluorescence images," *Sensors*, vol. 20, no. 23, p. 6713, 2020.
- [36] H. Pan, Z. Pang, Y. Wang, Y. Wang, and L. Chen, "A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects," *IEEE Access*, vol. 8, pp. 119 951–119 960, 2020.

# Augmented Reality: Prototype for the Teaching-Learning Process in Peru

Shalóm Adonai Huaraz Morales<sup>1</sup>  
Faculty of Sciences and Engineering  
Universidad de Ciencias y Humanidades, Lima, Perú

Laberiano Andrade-Arenas<sup>2</sup>  
Faculty of Sciences and Engineering  
Universidad de Ciencias y Humanidades, Lima, Perú

Alexi Delgado<sup>3</sup>  
Mining Engineering Section  
Pontificia Universidad Católica del Perú, Lima, Perú

Enrique Lee Huamani<sup>4</sup>  
Image Processing Research Laboratory  
Universidad de Ciencias y Humanidades, Lima, Perú

**Abstract**—In recent years, augmented reality is playing an important role in the world of mobile technology, since it is a way to facilitate the teaching-learning processes, this easy teaching-learning process generates a great contribution in companies, either creating opportunities or changing the way in which companies approach and interact with their end customers, this can mean a remarkable growth of the organization, that is why in this work an augmented reality prototype has been made using the methodology Scrum at the University of Sciences and Humanities of Lima-Peru, but with a focus on the nursing career. Knowing that the problem is the limited learning that students acquire in the classrooms, for which, we want to make use of augmented reality, so that this improves the form of education that is provided to university students. The result obtained, from developing the case study, was an augmented reality prototype for the improvement of education at the University of Sciences and Humanities of Lima-Peru, which shows a virtual model (it depends on the image shown), able to interact with the user, making it attractive and motivating for the student, this prototype was achieved, using Unity (3D development platform, Vuforia (augmented reality software development kit), Microsoft Visual Studio (integrated development environment), the Scrum Methodology (Scrum Pillars, Product Backlog, Product Backlog Estimation, Speed, Backlog Prioritization and Sprint Planning) and the C# Language (C Sharp).

**Keywords**—Augmented reality; teaching; education; scrum; unity

## I. INTRODUCTION

The use of the increased reality, in the world, is one of the most modern technologies, because it is considered to be a technique that offers privileges in education, giving a high potential to this, making the learning more attractive to the students. According to the [1] that is carried out, using the increased reality, the project offers the opportunity to work with the students, to motivate them and administer them knowledge through the improvement of the teaching of the real world, with a series of virtual objects within this, creating this year, learning experiences. According to this, [2], communications between teachers and students, which, when performing activities together, serves teachers to study as students acquire knowledge in the classroom, but only, providing them with a limited vision of learning, so this study, makes a useful contribution to existing knowledge.

Likewise, in research [3], about learning, it was more likely more than 70% of students identified learning experience, as a preferred method of learning. And the research [4] has shown how it is better learned and participating in "real" tasks in practice, instead of teaching theory, this can be developed through the teaching of cognitive processes such as perception, attention, memory, and thought. This is the case, that learning [5] experience requires a repetitive process, whereby ideas are applied and tested as feedback is received for improvement. According to China [6], reports in its studies on the use of increased reality tools to develop the oral competition of students in language learning, where data was collected, recorded oral presentations, observations, and comments from students.

These results revealed, which, with the use of the augmented reality tool, obtained a higher score, which when they have not augmented reality tools, these increased reality tools, showed students a facility in learning. In Peru [7], There is a problem with education, that is why the incorporation of augmented reality is an alternative for the improvement of some concepts, in the way that teachers teach students, to generate, susceptible to use new technologies in the field of teaching, giving positive results, since the increased reality, as a complement to teaching, will be beneficial. Therefore, the importance of this work lies in the fact that, slowly, but steadily, they allow augmented reality to be established among education professionals as a strategy that can transform the education of students.

All this was done through the use of the Scrum methodology, where the Pillars of Scrum were used so that the methodology is effective at the time of implementation, the Product Backlog where the user stories were placed, Estimating the Product Backlog where each story was scored of user, Velocity where it was established how many sprints were carried out, Backlog Prioritization where the user stories were ordered by priority and Sprint Planning where the development of each sprint was planned to develop the augmented reality prototype.

The objective of this work is to make a prototype of augmented reality under the Scrum methodology to carry out a pilot with the Nursing career.

Section II explains the literature review, Section III the methodology, Section IV the results and discussions, and

finally Section V explains the conclusions and future work.

## II. LITERATURE REVIEW

This section explains the fundamental material for the preparation of this paper.

### A. Background

A survey on augmented reality is approached, it tells us about the progress that augmented reality has had, showing us that augmented reality is compatible with many technologies [8]. It also shows us successful designs and applications of augmented reality in education, architecture and marketing aspects, in addition to dealing with important topics based on augmented reality such as augmented reality screen technologies (sensory screens), augmented reality development tools (Software frameworks, development tools and creation tools), augmented reality interaction technologies (Browsers and interfaces), interface patterns and evaluations of virtual reality systems (Methods).

This article [9] provides us with a wide range of various software development kits (SDKs) related to augmented reality, such as Metaio, Vuforia, D' Fusion, Wikitude, ARToolKit and AR-media; In order to compare them (by license, by platform, by markers and by the ability to superimpose) according to the qualities that each one of them has, it also shows us the differences between virtual reality and augmented reality, where its details are detailed. more relevant differences so that the reader can differentiate them quickly, in addition this article emphasizes the fact that mixed reality is the sum of augmented reality with virtual reality, and finally shows us how augmented reality works, teaching it in a bold way so that in this way its operation can be fully understood.

As for learning, this article [10] presents augmented reality focused on education, says augmented reality can be found in different areas, but in itself it will be more focused on education, provoked and thus generating unique environments; it also shows the uses, advantages, characteristics and effectiveness of augmented reality; He tells us about the analysis they made of augmented reality, following classifications such as countries, subject matter, type of augmented reality and research methods. To conclude, it infers us that there are numerous educational applications regarding augmented reality and that they are used in classrooms as well as abroad.

Another field in medicine, this article [11] mentions augmented reality but applied to surgery, it mentions that augmented reality can give efficiency and safety in surgical training; It also shows a scheme with basic principles of augmented reality as a basis for the reader to understand clearly and concisely "What is augmented reality?", this article mentions that the augmented reality system provides data in real time to the surgeon, Emphasizing the projection of augmented reality that can be given through the screen of computers, tablets, projectors, cameras and smart glasses, the article ends by making it clear that augmented reality is capable of revolutionizing the field of surgery.

To finish the background of the literature review section, the following article [12] will be addressed based on a survey on the applications of augmented reality, this article gives us

to understand that augmented reality is capable of improving human perception by mixing objects physical of artificial ones; It also gives us an apex of the initial applications that augmented reality addressed, these being medicine, manufacturing and repair, annotation and visualization, robot route planning, military applications and finally entertainment, as well as those initial applications Over time applications such as tourism, architecture, cultural heritage (Museums) and education were added; To finish this article, mention that augmented reality is related to miniaturization.

### B. Related Work

In this article [13], the use of mixed reality for the development of an Augmented Paper Map (APM) system was observed, the two authors who developed this article report Scrum, as the methodology that helped them to achieve the objectives of their article, explaining a bit about the roles and artifacts that were useful for the application of this methodology, they also express the design philosophy they used, which is that the design is user-centered, so that in this way the final product complies with the needs of these (Users), and thus achieve the greatest satisfaction with the minimum effort on the part of the user. It also tells us about the activities that they had to develop and carry out during the course of the article, as well as the exploration, analysis and evaluations that they had to develop for the development of the Augmented Paper Map (APM) system.

The mention of games is observed, it tells us drastically how the game is encompassed in the lives of human beings, it also mentions mixed reality but in the approach of games [14], it tells us like the previous article the use of Scrum , but focusing on the MVCE architecture (Model-View-Controller-Environment), which is nothing more than the MVC architecture (Model-View-Controller), but adding an "Environment" component. It also highlights the design of a mixed reality, but based on the independence of the device, the adaptive presentations and the updates of the context with respect to the game, as well as its philosophy of design centered on the user, which achieves a better experience for the game. (User) only with minimal effort.

As the clear examples of the application of the Scrum methodology, now there is a combination of the Scrum methodology with the extreme programming methodology (XP), which is focused on virtual reality. This article [15] shows us the use of a video game engine known as Unity, the Scrum plus XP combination to use it as a methodology as mentioned before and the objective of the article, which is the creation of virtual reality for the training of technicians and industrial operators. They focus on using virtual reality as an economical alternative to using and damaging expensive industrial equipment, so their objective is focused on two scenarios: instrument recognition activity (HART Device) and training activity (Installation, connection, basic and advanced settings).

Virtual reality focused on agile development, this article [16] shows the explanation of the use of the Scrum methodology, as well as the explanation of the use of the extreme programming methodology (XP), the principles of the agile manifesto are also covered, the life cycle of Scrum and extreme

programming (XP), the systems developed based on virtual reality, the application of agile methodologies in virtual reality and to finish they propose a process of development of virtual reality systems based on eight activities main (Definition of user stories, architectural peak, elucidation of interactivity requirements, iteration planning, peak, development, integration tests and customer tests).

### III. METHODOLOGY

In this section, the methodology that was carried out was explained, in the augmented reality prototype using Unity, the Scrum methodology was carried out, for the improvement of education at the University of Sciences and Humanities in Lima-Peru, thus achieving the objectives, with support of this methodology [17] that is shown in Fig. 1.

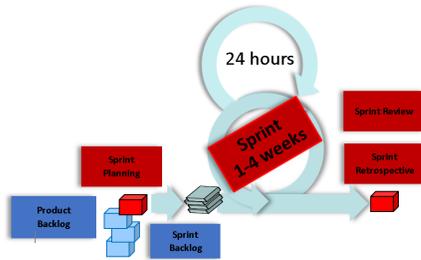


Fig. 1. Scrum Agile Methodology for the Augmented Reality Prototype.

As mentioned the prototype was made in Unity, which is a 3D development platform, to develop a variety of simulations, in games (developers of games), AEC (architecture, engineering, and construction), cars, and cinematographies [18]. And it is used, Vuforia (augmented reality software development kit), which is, one of the most popular to introduce the increased reality, Vuforia (augmented reality software development kit) uses vision technology to recognize and track images in real-time.

The increased reality based on Vuforia (augmented reality software development kit), takes the display of the device screen, as a "medium" to connect it to the world of the increased reality. What it does, is to show the images of the real world in the Camera of the corresponding device, adding the virtual objects 3D, in the images of the real world, what it does, that is a combination, of a feeling of immersion towards the world of the increased reality, also, to develop the programming in the prototype will be used the Microsoft Visual Studio integrated development environment, using programming language C# (C Sharp) [19].

#### A. Scrum

Scrum won't tell you exactly what to do, what it means, that can be done differently, as your merits, that is why the "power" is in us, to adapt it to a specific situation, and is here, where everything starts.

1) *Pillars of Scrum:* In the Fig. 2 the three pillars (transparency, inspection, and adaptation) that sustains SCRUM, the first pillar, the transparency, gives a clear vision to all the interested parties of the project, customers, users, sponsors, investor, among others. It tells us that clear agreements must

be reached on what must be delivered or informed to the interested parties, being thus, tracing and sincere, of "how much has been advanced?", "What is it will be achieved?" and "What will not be achieved?". Inspection, the second pillar, is where the project checks, to know if the established goals are being fulfilled and if there are problems or any deviation, to be able to correct them at the time.

In Scrum there are meetings such as the diary scrum (it is done when starting the day), or the retrospective (checking at the end of each sprint), this is done to know "What has been done?", "What problems have been?" and "What is going to do today?" That's why it can be said that in Scrum if there is a "constant inspection". Adaptation, the third pillar, means those that change is welcome, and this is so, to be able to minimize the problems that are generated in the future so that in this way, continue to provide the maximum value to the client, to the business, and/or to what is developing [20].

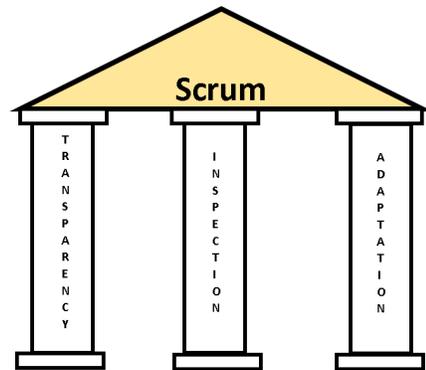


Fig. 2. The Three Pillars of Scrum.

2) *Project Charter:* Table I shows the Project Charter, which describes the objectives to be achieved in the project, it is here, where it is defined that the agile methodology is used., specifying the acceptance criteria for this one. In this Project Charter, the following questions are answered (Who?, What?, Where?, When?, Why? and how?), And is it here, where you can use the tool known as "Elevator Pitch", to perform the Project Charter [21].

TABLE I. PROJECT CHARTER

| For             | Target customer                   |
|-----------------|-----------------------------------|
| Who             | Need (Opportunity or problem)     |
| He              | Product / Service name            |
| What is it      | Product Category                  |
| For             | Key benefits / Reasons to buy it  |
| Is not equal to | Main competencies or alternatives |
| We              | Definition                        |

3) *Product Backlog:* Fig. 3 shows us the backlog prioritized, this is achieved, when performing meetings with the owner of the product, and thus generate the stories of users and prioritized (functionalities, epic, user stories), in addition to focusing on business terms, you are to be built in a short time, conducting valuable features for customers [22].

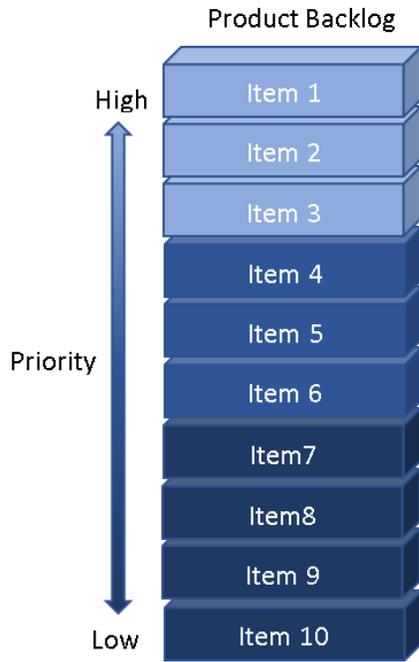


Fig. 3. Product Backlog.

4) *Estimating the Product Backlog:* Fig. 4 shows us how it would be an estimate of the Product Backlog, indicating its story points (sp), stories, must be reasonable, and these estimates can be generated with various tools, such as the size of your shirts, analogous estimate, Delphi and/or Planning Poker [23].

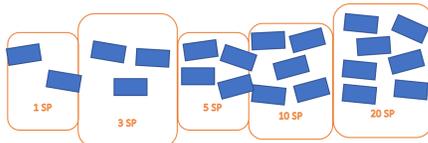


Fig. 4. Estimating the Product Backlog.

5) *Velocity:* In Scrum the speed is used, to define how many Sprints have, and this is, in function to speed, which is the stories of stories (necessary effort, with which the user stories were estimated) that a development team performs in each sprint. This speed may vary in each sprint, as shown in Fig. 5, wherein the chart of columns grouped in 3D, observed that the speed varies or changes by Sprint [24].

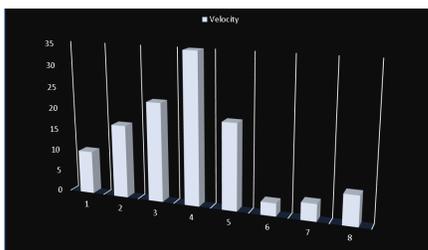


Fig. 5. Estimate Sprint Velocity.

6) *Backlog Prioritization:* In Fig. 6, the map of the story is displayed, which serves to prioritize the Product Backlog, the structure of this is the same as that of the Road Map, where its columns are the software or project modules, and where the first row is the backbone (most importantly, if this is not done. No system), the second row is the walking skeleton (Minimum viable product, which means, the most valuable functionality for the user, and the one that will give it to the system), and the following rows range from the most important user stories, at least important [25].

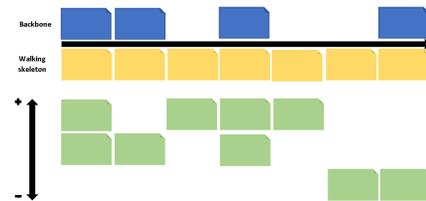


Fig. 6. Map of the Story.

7) *Sprint Planning:* In Fig. 7, it is observed that the structure of the Road Map [26], is the same as that of the map of the story, this is the map of the story sorted by delivery (Sprint set) or by Sprint (1 to 4 weeks), according to one establishes it, and is used to plan the Sprint. After performing the sprint planning, the user stories of the Sprint are selected to work, you are user stories will be developed, in a format of user stories, already established by the person in charge, as shown in Fig. 8, which has fields such as, the developer, the estimated time, the description of the user story, tests, tasks and finally the prototype; this format may vary [27].

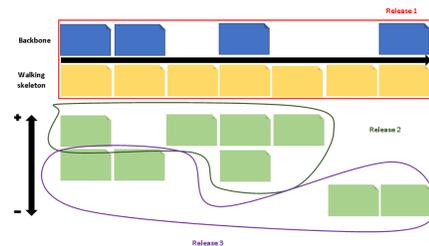


Fig. 7. Road Map.

| USER STORY           |               |
|----------------------|---------------|
| Number:              | Story name:   |
| User:                | Estimate:     |
| Developer:           |               |
| Description:         |               |
| Acceptance criteria: |               |
| Tests:               | Define tasks: |
| Prototype            |               |
|                      |               |

Fig. 8. User Story Format.

## B. Tools and Programming Language to Develop the Prototype

In this section, it was detailed, the tools to be used to develop the prototype of increased reality focused on improving teaching at the Universidad de Ciencias y Humanidades, as well as the programming languages to be used.

1) *Unity*: Unity is a 3D development platform, which gives us everything we need to be able to develop a variety of functionalities, such as simulating the laws of physics, adding animations, adding sounds and programming; This platform supports the C# (C sharp) programming language, in addition to having an Assets (resource) library known as the Asset Store of Unity, where you can find textures, animations and models ready to use in Unity [28].

2) *Vuforia*: To introduce the augmented reality, Vuforia (augmented reality software development kit) was used, since this uses vision technology to recognize and track the images in real-time, taking the display of the screen of the devices as a "medium" to connect it with the world of the increased reality. Vuforia (augmented reality software development kit). It serves us to generate a license key, which will help us load the database created in Vuforia (augmented reality software development kit) in Unity, to generate the augmented reality [29].

3) *Microsoft Visual Studio*: It is an integrated development environment (IDE) that supports several programming languages (C++, C#, Visual Basic .Net, F#, Java, Python, Ruby and PHP), this IDE gives developers the ability to create websites (Collection of web pages); Knowing that this IDE includes the C# (C Sharp) programming language, this is beneficial for this article, since the C# (C Sharp) programming language was used for the development of the augmented reality prototype [30].

4) *C#*: For the development of this augmented reality prototype, the programming language C# (C Sharp) was used, this programming language C# (C Sharp) can be used in several integrated development environments (IDE), such as Microsoft Visual Studio, SharpDevelop, QuickSharp, xacc.ide, MonoDevelop, and Xamarin Studio (XS); Since this programming language is compatible with Unity, this was the one that was used to give interactions of the "models" with the user, to achieve a kind of "communication" between the prototype and the user [31].

## C. Development of the Methodology

According to the methodology mentioned above, and the processes that were detailed in it, the development of the prototype will be implemented, based on the increased reality for the teaching of the courses of certain professional schools of the Universidad de Ciencias y Humanidades.

1) *Product Backlog*: In this section can observe the user stories that were developed to develop the prototype:

- User Story 1: As the director, I want to visualize a model that is related to a university course to be able to better explain the class.
- User Story 2: As the director I want to visualize multimedia content that is related to a university course to convey knowledge to students.

- User Story 3: As the director I want the prototype to be used in smartphones to have quick access to it.
- User Story 4: As the director I want the prototype to be interactive with the user to improve the education process.

2) *Estimating the Product Backlog*: To estimate the stories of users who were placed on the product stack (Product Backlog), a technique known as planning poker is going to be used, but not the planning poker is known as such, but a variant of it is, which works with the Fibonacci Sequence, which will be explained below. The Planning Poker has about eight "cards" (1/2, 1, 2, 3, 5, 6, 7,  $\infty$ ), without counting the "card" of doubt and rest, while the variance of the planning poker that goes from hand with the Fibonacci Sequence works with ten "cards" (0, 1/2, 1, 2, 3, 5, 8, 13, 21,  $\infty$ ), which will help us give you a "score", to each user story. The user's stories were beginning to be story through story (necessary effort), using the variation of Planning Poker, who works with the Fibonacci Sequence, was taken as a reference The user story that it is less than to make, making a "score" of "2" story points (necessary effort) and thus continue to find the estimates of the remaining user stories.

3) *Determination of Speed*: Here it is defined, how many sprints will be done, for this, first a "10" speed was determined, without forgetting, that each Sprint has user stories and these are in turn, they have "activities", and these "activities" are expressed in time (hours), which cannot exceed 4 weeks, since that deals with the agile methodologies, which each Sprint is completed from 1 to 4 weeks. After that, the sum of the story points was divided (necessary effort) of the user stories that were placed in the Product Backlog, between the certain speed which originated us "2.8", but when rounding it gives us "3", which means that for the prototype of this work have "3 Sprint". At the beginning of the first sprint, it was started with a speed of "10", which has a sum of "13" story points (necessary effort), and then, as it was advanced in the Sprint, this speed was changing, reaching the second Sprint, with a speed of "7", and with a sum of points of story (necessary effort) of "7", already in the last of the Sprint, the third Sprint, the speed was maintained in "7", and with a total amount of story points (necessary effort) " 8 ".

4) *Prioritizing the Backlog*: In this section, the backlog is prioritized, depending on the value given to the business, this is where, use the tool known as the map of the story, to generate the prioritization of the backlog. This map of the story, gives us an overview of "What is to do first?", and "What is to do after?". In the development of the map of story, it began, with the backbone that is the first "row" of the map of story, where it was placed, the essential user stories (as the director I want to visualize a model that is related to a course of the university to be able to better explain the class, as the director I want to visualize a multimedia content that is related to a course of the university to transmit knowledge to the students), to give improvement to education, then on the second "row", that of the walking skeleton was placed, the user story that gives the most valuable functionality for the user (as the director I want the prototype to be interactive with the user to improve the process of education), to end this section, in the last "row" remained the remaining user stories, ordered more important,

unless important (as the director I want the prototype to be used in smartphones to have quick access to it).

5) *Sprint Planning*: To plan the Sprint, the technique known as road map was used, as mentioned, in the methodology of this work, could be said that the path of the product, is originated from the map of the story. To start with this, the map of the story is taken, and starts to choose from, that user stories that was planned by Sprint.

6) *Sprint Backlog*:

- Sprint 1: In the first Sprint, as mentioned in the section of the speed determination, it was started with a "10" speed, making the user stories corresponding to this (as the director I want to visualize a model that is related to a course of the university to be able to better explain the class), for which, Unity is used, which is a 3D development platform, and Vuforia (augmented reality software development kit), to introduce the augmented reality, to which you ensemble the renovated reality, since this, uses vision technology to recognize and track the images in real-time, taking the visualization of the screen of the devices as "Medium" to connect it with the world of the increased reality, finally for this Sprint used web platforms used as free3d, to find and download 3D models, and find and download 3D models. In the first Sprint, in summary, what was done, was that the virtual 3 "objects" were added, to the 3D development unit platform, modifying the "object", since some of the components of the "object", not use it, it should be highlighted, that this modification was done in Unity, and the result of this, can observe it in Fig. 9.

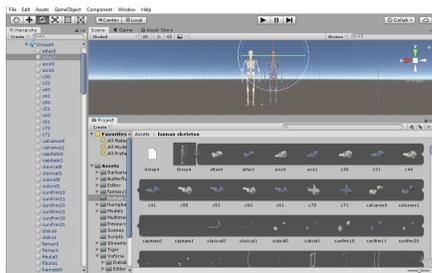


Fig. 9. Sprint 1 - User Story 1.

- Sprint 2: In the second Sprint, as mentioned in the section of the determination of the speed, it was begun with a "7" speed, making the user stories corresponding to this (as director I want to visualize a multimedia content that is related to a course of the university to convey knowledge to students, as director I want the prototype to be used in smartphones to have quick access to it), for which, Unity, Vuforia (augmented reality software development kit) was used and videos. To start with the second Sprint, in summary, what was done, was that the videos were added to 3D objects that can be created in Unity, the result of this can be seen, in Fig. 10, also, Unity (3D development platform) was used, so that the prototype can be used in smartphones, as can observe, in Fig. 11.



Fig. 10. Sprint 2 - User Story 2.



Fig. 11. Sprint 2 - User Story 3.

- Sprint 3: In the third and last sprint, as mentioned in the section of the determination of the speed, it remained with a speed of "7", making the user stories corresponding to this (as the director I want the prototype to be interactive with the user to improve the process of education), for which, Unity, Vuforia (augmented reality software development kit) was used, videos and the Microsoft Visual Studio integrated development environment. In the third, and last sprint, in summary, what was done, was that the Microsoft Visual Studio development environment was used to program, using programming language C# (C sharp), added this, to videos and models, to achieve that both the video and the model are interactive with the user, the result of this, can observe it in Fig. 12.

7) *Sprint Review*: This meeting was made at the end of each Sprint, to check how the objectives are being developed, it is meeting one hour by Sprint. Here what was done is to review the progress to reach the prototype, identified what was achieved, as well as what was not achieved, and is here where the operation of the end is shown in the Sprint.

8) *Sprint Retrospective*: This meeting, was made to give improvements to the following Sprint as it was moving forward, it should be noted, that this meeting was held after the Sprint Review with a duration of an hour, and is here where some questions were made ("worked on the last Sprint?", "What will be improved in the following Sprint?" and "What problems have been had in the progress of the development of the prototype?"), and at the end of the recommendations, the improvement was given to the following Sprint.

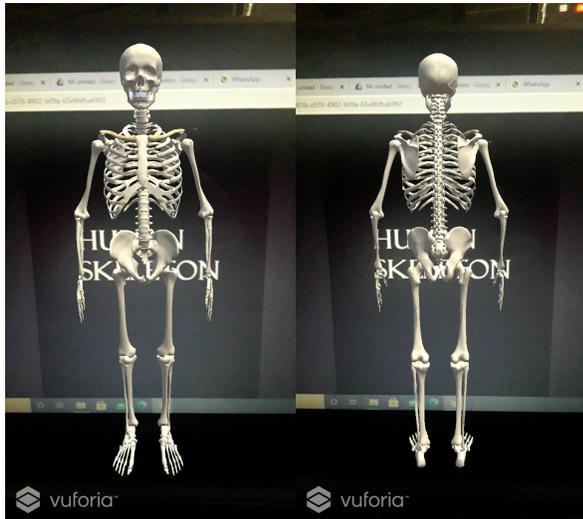


Fig. 12. Sprint 3 - User Story 4.

#### D. Testing

This section shows the survey administration software known as Google Forms, which can be seen in Fig. 13, where a survey was carried out among the students of the Faculty of Health Sciences of the Professional School of Nursing, in the It shows four short answer questions which are "What is your opinion about virtual reality?", "Would you like to visualize through the Smartphone 'objects' in 3D dimension?", "Are you likely to use this application?" and "Does its implementation seem correct?"; as well as a box question which is "For what course would you like this application?" This survey is key to the objective proposed in this paper, since the implementation of the prototype is focused on improving the teaching of the University of Sciences and Humanities, therefore, the opinion of the students cannot go unnoticed, since These are the ones that will handle the application, since this "application" is still a prototype, this does not evade the importance of the student's perception of virtual reality, because in the end it is for the student to feel satisfied with this implementation versus education, and thus goal-focused improvement is achieved.

Here is the before and after obtained from the Human Anatomy course, as shown in Table II, in the question "From traditional to automated", "From normal teaching to remote teaching" and "Contribution to the nursing career".

TABLE II. BASE LINE (BEFORE - AFTER)

| Before human anatomy course                | After human anatomy course                                            |
|--------------------------------------------|-----------------------------------------------------------------------|
| Traditional                                | Automated                                                             |
| Normal Teaching                            | Remote Teaching                                                       |
| Contribute to the nursing career community | Contribute to the nursing career community in these times of pandemic |

#### IV. RESULT AND DISCUSSION

In this section, the results of the case study should be shown, such as that of the Scrum methodology, used in this work.

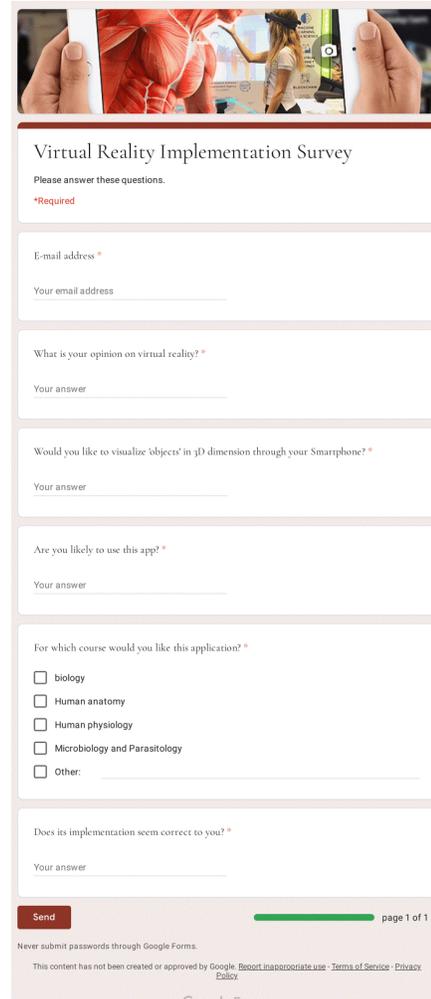


Fig. 13. Student Survey.

#### A. About the Survey

The survey carried out previously establishes specific questions, which can be observed in this subsection, but with the difference that the results obtained from each question asked in the survey are shown here, as shown in Table III, in the question "What is your opinion on virtual reality?" can observe a 48.1% knowledge about virtual reality by nursing students, while a higher number still do not know its meaning, in the question "Would you like to visualize 'objects' in 3D dimension through your Smartphone?" a low percentage of "37.8%" is established in the question "Are you likely to use this app?" A high interest in the application is observed since they find it curious to see how this prototype would perform in education, in "For which course would you like this application?", have varied percentages, since it is in a box question, these Results are respectively set to the boxes of question 4 shown in Fig. 13 and to finish a large number of students are observed who consider the implementation of this prototype correct.

The results of a survey carried out in the organization towards the General Manager, Coordinators, School Directors and the Dean are also observed, as shown in Table IV, in

TABLE III. SURVEY RESPONSE

| Questions                                                                      | Answers                                 |
|--------------------------------------------------------------------------------|-----------------------------------------|
| What is your opinion on virtual reality?                                       | 48.1 %                                  |
| Would you like to visualize 'objects' in 3D dimension through your Smartphone? | 37.8 %                                  |
| Are you likely to use this app?                                                | 89.9 %                                  |
| For which course would you like this application?                              | 95.7 %, 85.4 %, 80.2 %, 86.5 % y 86.1 % |
| Does its implementation seem correct to you?                                   | 94.3 %                                  |

the question "Are you satisfied with the scrum methodology carried out in the prototype?", Where a 65% acceptance is observed; in the question "Would you recommend using the scrum methodology for other projects?", where a 75% acceptance is observed; and to finish in the question "Do you think that Scrum methodology is better than the traditional methodology used in the university?", an 80% acceptance is observed.

TABLE IV. ORGANIZATION SURVEY

| Questions                                                                                              | Answers |
|--------------------------------------------------------------------------------------------------------|---------|
| Are you satisfied with the scrum methodology carried out in the prototype?                             | 65 %    |
| Would you recommend using the scrum methodology for other projects?                                    | 75 %    |
| Do you think that Scrum methodology is better than the traditional methodology used in the university? | 80 %    |

### B. About the Case Study

In the case of study, which deals with the development of the prototype of augmented reality, for the improvement of the teaching of certain courses of the professional schools of the University of Sciences and Humanities of Lima-Peru, was developed based on the Scrum methodology [32], which helped, on the process of development of the prototype as a "guide", since we showed the way for which it should advance for this development, this path can be observed and read in section 3 (case study) of this work, the 3D development platform was also used, the Unity [33], this being a good platform for the proposed in this work since the prototype was developed, thanks to this platform, as shown in Fig. 9, Fig. 10, Fig. 11 and Fig. 12 of point 6 (Sprint Backlog) of the SCRUM subsection belonging to the case of study, where it is observed, that in each of the Sprint it was vital, use the 3D replacement platform, for which, it could be said, that this platform is the center, of everything that was used, for the goal of this work, it should be noted that this 3D development platform, Unity, has a shop known as "Asset Store of Unity" to download models, textures and animations, and thus facilitate the process of development of the prototype, since if we find the model we require, it can be imported, and use it at the time, another platform, which was used, were the web platforms such as free3d, to find and download 3D

models, it was used, because in this, it is easier to find models that are required, since unlike "Asset Store of Unity", this is organized by categories (architecture, vehicles, characters, aircraft, furniture, electronics, animals, plants, weapons, sports, food and anatomy), it should be noted that as it is a model already created, Unity was used to modify it to the convenience of the prototype, as shown in Fig. 9, where a model containing two human skeletons, that of the left of an approximate color to the Black Squeeze is observed, which resembles the color of the bone, which we will need and the one of the right of the grayish blue color, which we will not need, so, we have to modify in Unity, this model that contains two human skeletons, also for the development of the prototype, was used, Vuforia (augmented reality software development kit) [34], to introduce the augmented reality, as it was mentioned, in the development of the study case, this is perfect for recognizing and tracking the images in real time, taking the display as "medium", to connect it with the world of the augmented reality, this can be seen in Fig. 12, where Vuforia (augmented reality software development kit) is doing their work, by recognizing and tracking in real time, the "Human Skeleton" image (behind the model), and connect it with the model of human skeleton, thus generating the augmented reality, which is key to this prototype, since the "essence" of this work, is the creation of a prototype of augmented reality, Vuforia (augmented reality software development kit) was used, since its handling is easy, apart from providing us with a database, where the images that will be tracked, to generate the augmented reality, and finally, for the development of the prototype, the Microsoft Visual Studio integrated development environment was used, in order to be programmed in C# language (C sharp), since it was wanted, that the prototype has interactions with the user. All mentioned in this subsection of "Results and Discussions", was made, to achieve good control and result in the development of the prototype.

### C. About the Methodology

The method of scrum [35], was very helpful, as we guide us, in the development of the prototype, to start, helped us to focus requirements, thanks to user stories, also, to estimate these stories, to know what the user's stories that are found to make them more, as those that were easier to develop, they also helped us, namely what amount of effort necessary we could do by Sprint, this became known, thanks to the determination of the speed; And thanks to the prioritization part of this methodology, where the story map was used, it was possible to know which user stories are more important for the prototype and which stories are less important, on the other hand, the path or route of the product was also used to know how much sprint, and which user stories are going to be done by sprint, and at the end of each sprint was made a review to see if the objectives were achieved, as well as, At the end of each sprint review, the feedback was used for continuous improvement.

- Benefits: The benefits offered by the Scrum methodology are varied [36], since when deciding to make the prototype using an agile methodology, it allows us to have a finished "product" in less time than when choosing a traditional methodology, that's why, here, we are dealing with specific things, which strengthened the development of the prototype, that's why we

will start with the speed, which made us know, how much effort was invested per sprint, but this, depending on the performance, that is giving the sprint. Also, thanks to the Scrum methodology, the development of the prototype could be divided into parts, thus generating a "finished product" by sprint. On the other hand, another benefits of the scrum methodology is the sprint review, as this helps us to verify the achievement of the goal, identifying what was achieved, as well as what was not achieved, and finally, the feedback sprint, offered by the scrum methodology, is good, as this allows us to make improvements towards the following sprint.

- Comparison: If compare the scrum methodology, with the methodology RUP (rational unified process) [37], can say, that the methodology RUP (rational unified process), is rigid, meaning that if the customer wants a change, it is very difficult to make it, since you follow the plan until the end of the project development, whereas in the Scrum, this is possible, because, scrum is flexible to change, another comparison, you can make, The reason for this is that, in the methodology RUP (rational unified process), the project is conceived as one, while in scrum the project is divided into parts. It can also be said that in the methodology, the deliverable is at the end of the entire project development, while in scrum the delivery is constant [38]. This comparison of Scrum and RUP can be seen in Table V.

TABLE V. SCRUM VS TRADITIONAL METHODOLOGY RUP

| SCRUM                                      | RUP                                                        |
|--------------------------------------------|------------------------------------------------------------|
| Accept changes                             | Resists changes                                            |
| Its development is flexible                | Its development is rigid                                   |
| The client is part of the development team | Customer communicates with development team                |
| Client available throughout the project    | Client available at the beginning of the project           |
| Value is delivered to the customer early   | Value is delivered to the client at the end of the project |

## V. CONCLUSION AND FUTURE WORK

This augmented reality prototype has been developed satisfactorily, and thus this will help to improve the teaching in the Universidad de Ciencias y Humanidades, since this augmented reality prototype, by teaching through cognitive processes, is suitable for courses in biology, human anatomy, human physiology, microbiology, and parasitology, belonging to the professional school of nursing. The tools used for the development of this prototype were efficient and valuable, both in the use of the 3d development platform, in the ability to introduce augmented reality, and in the interactions of the "models" with the users. And the methodology used, the scrum, was very efficient, for the development of the prototype, being the scrum processes, key, to achieve the goal set, and thus ensure the development of the prototype. In future research, a menu system could be implemented for the prototype, which has options, so that the users, can know, how the prototype is used, which will show the scanned images, and thus, achieve a better understanding with the user.

## REFERENCES

[1] P. P. Nechypurenko, T. V. Starova, T. V. Selivanova, A. O. Tomilina, and A. Uchitel, "Use of augmented reality in chemistry education," in

*Proceedings of the 1st International Workshop on Augmented Reality in Education Kryvyi Rih, Ukraine, October 2, 2018*, no. 2257. CEUR Workshop Proceedings, 2018, pp. 15–23.

[2] J. Garzón, J. Pavón, and S. Baldiris, "Systematic review and meta-analysis of augmented reality in educational settings," *Virtual Reality*, vol. 23, no. 4, pp. 447–459, 2019.

[3] S. H. Halili, "Technological advancements in education 4.0," *The Online Journal of Distance Education and E-Learning*, vol. 7, no. 1, pp. 63–69, 2019.

[4] A. Syawaludin *et al.*, "Development of augmented reality-based interactive multimedia to improve critical thinking skills in science learning," *International Journal of Instruction*, vol. 12, no. 4, pp. 331–344, 2019.

[5] N. Elmqaddem, "Augmented reality and virtual reality in education. myth or reality?" *International journal of emerging technologies in learning*, vol. 14, no. 3, 2019.

[6] M.-P. Chen, L.-C. Wang, D. Zou, S.-Y. Lin, H. Xie, and C.-C. Tsai, "Effects of captions and english proficiency on learning effectiveness, motivation and attitude in augmented-reality-enhanced theme-based contextualized efl learning," *Computer Assisted Language Learning*, pp. 1–31, 2020.

[7] P. Chen, X. Liu, W. Cheng, and R. Huang, "A review of using augmented reality in education from 2011 to 2016," *Innovations in smart learning*, pp. 13–18, 2017.

[8] D. Chatzopoulos, C. Bermejo, Z. Huang, and P. Hui, "Mobile augmented reality survey: From where we are to where we go," *IEEE Access*, vol. 5, pp. 6917–6950, 2017.

[9] A. Hanafi, L. Elaachak, and M. Bouhorma, "A comparative study of augmented reality sdks to develop an educational application in chemical field," in *Proceedings of the 2nd International Conference on Networking, Information Systems & Security*, ser. NISS19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: <https://doi.org/10.1145/3320326.3320386>

[10] H. Altinpulluk, "Determining the trends of using augmented reality in education between 2006-2016," *Education and Information Technologies*, vol. 24, no. 2, pp. 1089–1114, 2019.

[11] P. Vávra, J. Roman, P. Zonča, P. Ihnát, M. Němec, J. Kumar, N. Habib, and A. El-Gendi, "Recent development of augmented reality in surgery: a review," *Journal of healthcare engineering*, vol. 2017, 2017.

[12] L. F. de Souza Cardoso, F. C. M. Q. Mariano, and E. R. Zorzal, "A survey of industrial augmented reality," *Computers & Industrial Engineering*, vol. 139, p. 106159, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S036083521930628X>

[13] G. Ameri, J. S. Baxter, D. Bainbridge, T. M. Peters, and E. C. Chen, "Mixed reality ultrasound guidance system: a case study in system development and a cautionary tale," *International journal of computer assisted radiology and surgery*, vol. 13, no. 4, pp. 495–505, 2018.

[14] M. Cowling, M. Hillier, and J. Birt, "Integrating mixed reality spatial learning analytics into secure electronic exams," in *International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education*. ASCILITE, 2018, pp. 330–334.

[15] E. Lozada-Martinez, J. E. Naranjo, C. A. Garcia, D. M. Soria, O. R. Toscano, and M. V. Garcia, "Scrum and extreme programming agile model approach for virtual training environment design," in *2019 IEEE Fourth Ecuador Technical Chapters Meeting (ETCM)*, 2019, pp. 1–5.

[16] T. Riemann, A. Kreß, L. Roth, S. Klipfel, J. Metternich, and P. Grell, "Agile implementation of virtual reality in learning factories," *Procedia Manufacturing*, vol. 45, pp. 1–6, 2020, learning Factories across the value chain – from innovation to service – The 10th Conference on Learning Factories 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2351978920310672>

[17] S. M. Saleh, S. M. Huq, and M. A. Rahman, "Comparative study within scrum, kanban, xp focused on their practices," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–6.

[18] A. Juliani, V. Berges, E. Vckay, Y. Gao, H. Henry, M. Mattar, and D. Lange, "Unity: A general platform for intelligent agents," *CoRR*, vol. abs/1809.02627, 2018. [Online]. Available: <http://arxiv.org/abs/1809.02627>

[19] F. Peng and J. Zhai, "A mobile augmented reality system for exhibition hall based on vuforia," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, 2017, pp. 1049–1052.

[20] A. Azanha, A. R. T. T. Argoud, J. B. de Camargo Junior, and P. D. Antonioli, "Agile project management with scrum: A case study of a

- brazilian pharmaceutical company it project,” *International Journal of Managing Projects in Business*, 2017.
- [21] D. Owens, J. W. Merhout, and D. Khazanchi, “Project management assurance in agile projects: Research in progress,” *MWAIS 2018 Proceedings*, pp. 17–18, 2018.
- [22] N. Bolloju, S. L. Alter, A. Gupta, S. Gupta, and S. Jain, “Improving scrum user stories and product backlog using work system snapshots,” in *AMCIS*, 2017.
- [23] M. Adnan and M. Afzal, “Ontology based multiagent effort estimation system for scrum agile method,” *IEEE Access*, vol. 5, pp. 25 993–26 005, 2017.
- [24] S. Semenkovich, K. O.I., and K. Degtiarev, “A modified scrum story points estimation method based on fuzzy logic approach,” *Proceedings of the Institute for System Programming of the RAS*, vol. 29, pp. 19–38, 01 2017.
- [25] Y. Wang, J. Ramadani, and S. Wagner, “An exploratory study on applying a scrum development process for safety-critical systems,” in *International Conference on Product-Focused Software Process Improvement*. Springer, 2017, pp. 324–340.
- [26] R. Paul and L. Behjat, “Using principles of scrum project management in an integrated design project,” in *Proceedings of the 15th International CDIO Conference*, 2019, pp. 716–729.
- [27] L. Gonçalves, “Scrum,” *Controlling & Management Review*, vol. 62, no. 4, pp. 40–42, 2018.
- [28] V. T. Nguyen and T. Dang, “Setting up virtual reality and augmented reality learning environment in unity,” in *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2017, pp. 315–320.
- [29] M. Sarosa, A. Chalim, S. Suhari, Z. Sari, and H. Hakim, “Developing augmented reality based application for character education using unity with vuforia sdk,” in *Journal of Physics: Conference Series*, vol. 1375, no. 1. IOP Publishing, 2019, p. 012035.
- [30] A. J. R. Desierto, A. S. A. Recaña, J. C. T. Arroyo, and A. J. P. Delima, “Goonar: A bilingual children storybook through augmented reality technology using unity with vuforia framework,” *International Journal*, vol. 9, no. 3, 2020.
- [31] O. Comber, R. Motschnig, H. Mayer, and D. Haselberger, “Engaging students in computer science education through game development with unity,” in *2019 IEEE Global Engineering Education Conference (EDUCON)*, 2019, pp. 199–205.
- [32] B. A. A. Ammourah and S. A. Pitchay, “Challenges of applying scrum model and knowledge management for software product management,” in *RITA 2018*. Springer, 2020, pp. 123–130.
- [33] S. Chapagain, “Application development with vuforia and unity3d,” 2018.
- [34] E. Cieza and D. Lujan, “Educational mobile application of augmented reality based on markers to improve the learning of vowel usage and numbers for children of a kindergarten in trujillo,” *Procedia Computer Science*, vol. 130, pp. 352–358, 2018. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050918304046>
- [35] A. Tupia-Astoray and L. Andrade-Arenas, “Implementation of an e-commerce system for the automation and improvement of commercial management at a business level,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120177>
- [36] A. Carrion-Silva, C. Diaz-Nunez, and L. Andrade-Arenas, “Admission exam web application prototype for blind people at the university of sciences and humanities,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 12, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0111246>
- [37] R. Arias-Marreros, K. Nalvarte-Dionisio, and L. Andrade-Arenas, “Design of a mobile application for the learning of people with down syndrome through interactive games,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0111187>
- [38] M. J. Mbuh, R. Mbih, and C. Wendi, “Water quality modeling and sensitivity analysis using water quality analysis simulation program (wasp) in the shenandoah river watershed,” *Physical Geography*, vol. 40, no. 2, pp. 127–148, 2019. [Online]. Available: <https://doi.org/10.1080/02723646.2018.1507339>

# On the Long Tail Products Recommendation using Tripartite Graph

Arlisa Yuliawati, Hamim Tohari, Rahmad Mahendra, Indra Budi  
Faculty of Computer Science  
Universitas Indonesia  
Depok, Indonesia

**Abstract**—The growth of the number of e-commerce users and the items being sold become both opportunities and challenges for e-commerce marketplaces. As the existence of the long-tail phenomenon, the marketplaces need to pay attention to the high number of rarely sold items. The failure to sell these products would be a threat for some B2C e-commerce companies that apply a non-consignment sale system because the products cannot be returned to the manufacturer. Thus, it is important for the marketplace to boost the promotion of long-tail products. The objective of this study is to adapt the graph-based technique to build the recommendation system for long-tail products. The set of products, customers, and categories are represented as nodes in the tripartite graph. The *Absorbing Time* and *Hitting Time* algorithms are employed together with the *Markov Random Walker* to traverse the nodes in the graph. We find that using *Absorbing Time* achieves better accuracy than the *Hitting Time* for recommending long-tail products.

**Keywords**—Long tail; recommender system; tripartite graph; random walker; hitting time; absorbing time

## I. INTRODUCTION

Promotion becomes one of the success factors in product marketing [1], i.e. the better the promotion, the more people recognize the products being promoted, and the higher chance for those products being sold. On the other hand, the wrong strategy in promoting products could also cause difficulty or even failure in selling specific products [2]. In general, e-commerce companies tend to recommend popular products to customers. Those products would remain popular and the not recommended products would be less exposed by customers. The high number of unpopular products would be harmful to some B2C (Business-to-Consumer) e-commerce companies that apply the non-consignment sale system. These e-commerce companies have already paid the products from the manufacturer to be sold to the customers. The products will not be returned even though the companies are unable to sell them. The more unpopular products are unsold, the higher the cost would be due to the damage risk or the inventory cost for storing such products.

The long-tail phenomenon is a condition when the unpopular (niche) products dominate the total sales [3]. Long-tail products are also interpreted as the less popular products among customers [4]. Even though the sales volume of each product was not so high, the total number of products dominate the total sales [2]. The ratio between long-tail products and popular products is following the 80/20 principle or Pareto rules. The 80% of total revenue is obtained from 20% of total products, i.e., the popular ones. By increasing the sales volume

of the remaining 80% of the total products (the long tail), the total revenue could be increased significantly.

A recommendation system is one of the important tools for marketing strategy. It is useful in dealing with the information overload issue as the variation of the products increases. Many studies in the recommendation systems for the e-commerce domain have been conducted [5], [6], [7], [8], [9]. Studies in this area are usually focused on the behavior or characteristics of the "known" products or "the shopping history" of customers. The common objective is to recommend the most suitable products based on transaction history. By its ability to capture customers' preferences, it is easier to recommend such suitable products for them. And for the customers, it will be easier to determine which to purchase and where to buy. On the other hand, since the recommendation commonly brings popular products up, these products become more competitive among many business owners [4]. Thus gaining profit from such products could be more challenging. On contrary, the less popular products are less noticed by many sellers so they could bring more profit if it is successfully sold [10], [4].

The characteristic of the data being used in a recommendation system is suitable to be represented in a graph. In every domain of the recommendation system, it is possible to represent the entities, such as users, items, movies, foods, images, books, as the nodes (vertices) of a graph. Meanwhile, any relations between entities can be represented as the edges. Many studies about graph-based recommendation systems have been conducted for different problems to be solved. A graph structure was owned by mostly recommendation system and also it raises many potential exploration and development through graph learning [11]. One common graph representation is a bipartite graph, for example, to capture the relationship between a set of users and a set of items. A study by [12] used this kind of representation to apply collaborative filtering based on user similarity and item similarity. A bipartite graph was also used by [2] to solve the long-tail problem through a random walker that is adopted in this study. The other example with different graph representation but also employed the random walker is a study by [13]. It solves the cold start problem through a trust network by applying trust-based and item-based collaborative filtering.

Now recalling the non-consignment sale system applied to some B2C e-commerce companies, besides capturing customers' preferences, it is also important to take the unpopular products out to the customers. The motivation behind this study is to find the recommended products, which not only focus on the more popular products but also those which

are less exposed by customers but still in the customers' preference area. Adopting previous studies, a tripartite graph representation is used to draw the relation between users, items, and categories. Since customers nodes are only connected to product nodes that they have ever purchased, then to make them exposed by the long-tail products, the *Markov Random Walker* combined with *Hitting Time* or *Absorbing Time* is employed to find the unpopular yet suitable products to be recommended to the users. In addition, as the product categories are available at different levels, this study also tries to figure out whether the different category level being used affects the recommendation results. More specifically, studies about the recommendation system for long-tail products are presented in Section 2 followed by the detail about tripartite graph implementation in Section 3. Section 4 consists of the experimental result and its analysis, and the conclusion would be presented in the last section.

## II. RELATED WORK

Studies to deal with the long-tail problem have been widely conducted. A study in [14] tries to analyze and solve the long-tail problems on the traditional recommendation system (i.e. collaborative filtering) on an e-commerce platform. By capturing the users' information and behavior together with the systems' behavior, several models are established involving Gradient Boosting Decision Tree, Logistic Regression, and user entropy-based LDA. This study shows that it is possible to recommend long-tail products while maintaining the quality of the recommendation.

A similar conclusion was also obtained from a study in [15] to enhance the collaborative filtering such that it considers mining the long-tail items in the recommendation process. This study was conducted on the sales of alcoholic beverages (RateBeer). A matrix factorization was established based on personal experience to generate the user experience level. The top  $N$  recommendation is then obtained from the experience level together with the consideration of items' popularity. This study captured a phenomenon where the customers with lower experience levels tend to purchase popular products more, and vice versa: those with higher experience levels, tend to purchase the unpopular ones. The problem of this study happened for new customers. The recommendation was either not relevant or only focused on the popular products.

Specific to deal with the cold-start problem and long-tail problem, a study by using social data (Flickr, BlogCatalog, YouTube, HetRec11-LastFM) conducted in [10] first decompose the overall products into the low-rank (short-head) products and the sparse part (long-tail) products. These two groups were trained independently and the final recommendation from each group was merged then became the recommended products for the new users. But basically, this study focuses more on resolving the cold start problem while 'introducing' some items from the long-tail category in the recommendation. Experimenting on a similar domain, especially related to movie viewers data (through MovieLens and Last.fm), a study by [16] developed CORE (Cosine Pattern-based Recommender). This system allows product recommendation based on either the popular products (based on those which have been rated by a user) or the niche products (based on the *cosine* pattern. This

study reported that the accuracy of the recommendation will be decreased when comes to dense data.

Other studies in recommendation system studies employ the graph representation. Specific for the long-tail problem in a movie data set, a study in [2] initiate the use of a bipartite graph to represent the user-item relation. *Markov Random Walker* was implemented to calculate the *Hitting Time*, *Absorbing Time*, and *Absorbing Cost* which were used to determine the ranking of the product recommendations. The performance of *Absorbing Cost* outperformed the other two on various measurement metrics used due to the characteristic of the *Absorbing Cost* that considers customer interests/preferences when giving a recommendation. This is suitable if all products are similar as in movie data because it is easier to recommend a movie to customers who have a specific interest in a particular genre compared to customers who have an interest in several genres.

Adopting the bipartite graph approach in [2], a study in [17] added latent information (i.e genre node) as a link between the customer node and the recommended product node such that the graph representation became a tripartite graph. By this improvement, it is possible to traverse from a user node  $x$  to the item node  $y$  that is not directly connected through the intermediate genre node  $z$  that might be indirectly connected to  $y$  (for example through the intermediate node). The result shows its ability to 'pick' the recommended items from the region that is suitable to the users' taste. A study in [4] makes an improvement in determining the latent information by using a single category. This study was also proposed a new approach to calculate the weight between product and category nodes to avoid the misleading caused by the use of direct average rating, namely the Bayesian averages. It shows the recall and the diversity score improvement compared to the former study in [17]. In both of these studies, the *Hitting Time* and *Absorbing Cost* was employed based on their performance in the study by [2].

Compared to the movie data (MovieLens) used by [17] and [4], the domain of this study (e-commerce data from B2C company) owns similar components. Both data own set of users, items (i.e. movies vs products), and category (i.e. genres vs products' categories) such that it is possible to build the tripartite graph representation. This study then adopts the approach used by [17] and [4] to build the tripartite graph-based recommendation system that employs the *random walker* to promote the long-tail item to the user. However, the characteristic of e-commerce data is different from the movie data, so this work differs from the former at some points. The first one is when customers often purchase products from certain categories than the other, it does not always imply that the category is preferable to the other. For example, because a customer often purchases snacks on e-commerce, this does not imply that the customer does not interested to purchase clothes or electronic devices. This is probably the customer prefers to buy clothes or electronic devices on offline stores rather than from e-commerce. This is surely different from movie data where the preference to watch movies for a specific genre generally implies the preference to the respective genre. From this condition, having information about what a customer often purchased from e-commerce is not too useful for the recommendation process, so that *Absorbing Cost* that works

by considering such information to make a recommendation is not suitable to be applied on e-commerce data. Thus, even though [2] mentioned that *Absorbing Cost* could recommend better than *Hitting Time* and *Absorbing Time*, this study try the other way –to not include the *Absorbing Cost*. The second point is that latent information in the movie data (genre) is very different from the latent information in e-commerce data (category). The main difference is that movie genres have the same level, while the product categories are divided into general categories (top-level) to the most specific category (leaf-level). Since the relation (as well as the weight) between items and different levels of the category might be different, the use of different levels of product categories could be one thing to be elaborated on, whether or not it affects the result of the recommendations.

### III. TRIPARTITE GRAPH RECOMMENDATION SYSTEM

#### A. Graph Representation

This study employs a tripartite graph, i.e. a graph  $G = \{V, E\}$  which its node set  $V$  is partitioned into three disjoint node subsets  $V_1, V_2$ , and  $V_3$  such that  $V = V_1 \cup V_2 \cup V_3$  and for each  $(u, v) \in E$ , if  $u \in V_i$  and  $v \in V_j$ , then  $i \neq j$  [18]. The graph representation utilized in this study would be an in-directed graph since a relationship between two nodes implies the reverse relationship. There are three types of nodes such as the *user* nodes representing the customers, *item* nodes representing the products, and *category* nodes representing the categories. The relationship between nodes is represented as weighted edges. There are also three types of edges connecting nodes from different types, they are the *user-item* edges, *item-category* edges, and the *user-category* edges.

Illustration of a tripartite graph representation is presented in Fig. 1. The blue, red, and green nodes represent the group of user or customer nodes, category nodes, and the item or product nodes, respectively. The label for user nodes was taken from the customers' username, while the product category label represents the product category name, and the item nodes use the brand name of the products as their label. In Fig. 1 each edge connecting nodes in the different groups represent different relationships. For example, the user node labeled "agustini24" has a pair of edges with opposite directions that are connected to the "Home Living" node. These edges represent that "agustini24" purchased products belonging to the "Home Living" category, and vice versa, the products belonging to the "Home Living" category were purchased by a user with the username "agustini24". Similar relations are also applied for the other edges connecting nodes from different groups of nodes. The rest of this subsection discusses detailed information about edge representation.

1) *User-Item Edges*: These edges connecting the *user* nodes with the *item* nodes. The weight of this edge is 1 if a customer has given a rating to a product, and 0 otherwise. To avoid the density of the graph, edges whose weight is 0 are removed.

2) *Item-Category Edges*: These edges connect the *item* nodes to *category* nodes. This type of edge uses the average rating value from all customers for a product in a specific category as the weight. It is computed from the average rating for a product  $i$  (denoted by  $i_{avg}$ ) divided by the total number

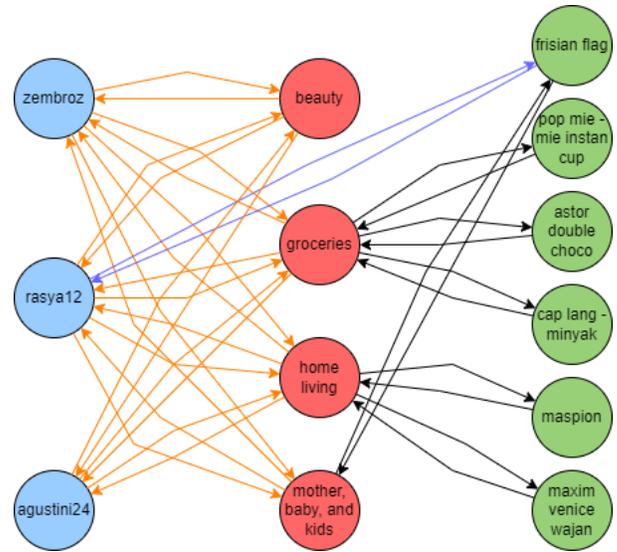


Fig. 1. An Illustration of the Tripartite Graph.

of categories connected to the product  $i$  (denoted by  $|C|$ ). The formula is presented as

$$w_{i,cat} = \frac{i_{avg}}{|C|} \quad (1)$$

3) *User-Category Edges*: These last group of edges exist between the *user* and *category* nodes. These node connect *user* node to more *item* node with shorter paths. The weight is computed by using *Bayesian Average*

$$w_{u,cat} = \frac{avg\_votes_u \times avg\_rating_u + votes_{u,cat} \times rating_{u,cat}}{avg\_votes_u + votes_{u,cat}} \quad (2)$$

for edges connecting *user* node  $u$  and *category* node  $cat$ . The value of  $avg\_votes_u$  is obtained by calculating the number of products bought by customer  $u$  in category  $cat$  divided by the total number of categories. The value of  $avg\_rating_u$  is calculated by dividing the total ratings from customer  $u$  for all products with the total number of categories. The  $votes_{u,cat}$  denotes the number of products in category  $cat$  that has been purchased by customers  $u$ . And the last,  $rating_{u,cat}$  denotes the average ratings from customer  $u$  in a category.

#### B. Product Recommendation

This section contains an explanation about how to get the recommended long-tail product for certain customers by traversing the tripartite graph. The long-tail products are determined based on the average number of customers that give ratings to the whole product in the data set. A product is then labeled as long-tail if the number of customers that give a rating to it is below the average.

1) *Markov Random Walker*: A *random walk* is formed from a graph traversal such that given a starting node  $a$ , we choose an adjacent node  $b$  to be visited at random (usually based on predefined transition probability), then choose the

next random node  $c$  to be visited from  $b$  and so on until certain steps [19], [11]. On a weighted graph, it forms a Markov Chain with the transition matrix consisting of the probability value of the movements between node  $i$  and  $j$  such that

$$p_{i,j} = \frac{w_{i,j}}{d_i} \quad (3)$$

where  $w_{i,j}$  denotes the weight between node  $i$  to node  $j$  while  $d_i$  is the total weight of node  $i$  to all of its adjacent nodes [20]. In this study,  $p_{i,j}$  refers to the probability of the *random walker* arrives at a product node  $j$  from a customer node  $i$  as the time  $t$  increases. The transition matrix (which is then denoted as  $M_S$ ) is also called as a *stochastic matrix* since the sums of each row equals to 1.

In the matrix representation of a graph, the dot product of a matrix by itself for  $n$  times results in the availability of paths with length  $n$  between each pair of nodes. Related to this study, from a given *user* node, an iterative process is done by the *random walker* to find the suitable long-tail *item* node. This process is equivalent to the dot product of the *stochastic matrix*  $M_S$  by itself for  $t$  times which represents a *random walker* probability traverse from node  $i$  to node  $j$  in time  $t (\geq 1)$ . As the time  $t$  increases, the elements in this *stochastic matrix* converge such that no change in their value or the changes are very small. However, according to [17], it is better to use a small value of  $t$  since as it grows higher, the *random walker* tends to visit the popular nodes. Thus in this study,  $t = 2, 3, \dots, 7$  are used since after  $t = 7$ , the ability of the system to recommend the long-tail products was decreased [4].

2) *Hitting Time*: As stated in [2], *Hitting Time* (denoted as  $H(q|j)$ ) is defined as the expected number of steps that is needed by a *random walker* to move from an *item* node  $j$  to *user* query node  $q$  with  $j \neq q$ . The value of *Hitting Time* is obtained from

$$H(q|j) = \frac{\pi_j}{p_{q,j} \cdot \pi_q} \quad (4)$$

where  $\pi_j$  and  $\pi_q$  are the *stationary probability* for node  $j$  and node  $q$  respectively. Meanwhile,  $p_{q,j}$  represents the weight of edge connecting node  $q$  and node  $j$ , i.e. the movement probability between node  $q$  and node  $j$ . The smaller the value of  $H(q|j)$  denotes the more relevant node  $q$  and  $j$  and that only a few users have rated item  $j$ . This conclusion comes from the following information.

- Consider the fact the value of *stationary probability* stays constant for all nodes, the value of the *Hitting Time* is inversely proportional to  $p_{q,j}$ . This means that the higher the value of  $p_{q,j}$  which denotes the more relevant node  $q$  and  $j$ , then the lower the *Hitting Time* value will be obtained.
- The *stationary probability* of a node is proportional to the number of customers that give a rating to the product. This means that the lower the *stationary probability* of a product, then it belongs to the long-tail product because it is only rated by a few customers.

---

**Algorithm 1.** Recommendation by using Hitting Time

---

**Input:**

A tripartite graph  $G = (V, E)$   
A customer node  $q \in V$   
Time  $t$  for how long the *random walker* traverse the nodes

**Recommendation\_By\_HT( $G, q, t$ ):**

- 1) define a subgraph  $G' = (V', E')$
  - 2) for each node  $j \in V$  that has not been rated by customer  $q$ :
  - 3) include node  $j$  as the member of  $V'$  in  $G'$
  - 4) include edge  $(q, j) \in E$  as the member of  $E'$  in  $G'$
  - 5) create a stochastic transition matrix  $M_S$  from  $G'$
  - 6) for each node  $j$  in subgraph  $G'$ :
  - 7) calculate the stationary probability ( $\pi_j$ )
  - 8) perform dot products:  $(M_S)^t$  represents the random walk length  $t$
  - 9) for each node  $j$  in subgraph  $G'$ :
  - 10) calculate the *Hitting Time* value  $H(q|j) = \frac{\pi_j}{p_{q,j} \pi_q}$
  - 11) sort the *Hitting Time* value for all node  $j$  in ascending order, except node  $q$
- 

Algorithm 1 presents the steps to recommend the unpopular products (i.e. the long-tail products) based on the *Hitting Time* value. This algorithm intuitively tries to find the product nodes which has never been purchased by a customer but have higher similarity to those that have been purchased by the customer.

3) *Absorbing Time*: As the comparison of *Hitting Time*, the *Absorbing Time* is implemented regarding the study by [2]. It explains that *Absorbing Time* is suitable for data in which the number of customer nodes is far higher than the product nodes. Within this condition, the number of average rating for each product is higher than the number of average rating for each customer. Thus this information should be more useful for the recommendation process. *Absorbing Time* that is denoted by  $AT(S|i)$  is defined as the expected number of steps before a *random walker* that is started from node  $i$  is absorbed by  $S$ . While the set  $S$  denotes the *Absorbing Nodes*, i.e. set of nodes  $S \subseteq V$  in a graph  $G = (V, E)$  for which the *random walker* stops when any node in  $S$  is reached for the first time.

$$AT(S|i) = \begin{cases} 0 & , i \in S \\ 1 + \sum_{j=1}^n p_{i,j} \cdot AT(S|j) & , i \notin S \end{cases} \quad (5)$$

calculates the value of *Absorbing Time*. It can be seen that  $AT(S|i)$  would be 0 whenever the current node is one of the *Absorbing Node*, i.e those which are directly connected to the customer node. While a recursive calculation is performed from the source node (the customer node) to one of the *Absorbing Node*. This approach is similar to the one that uses *Hitting Time*, to find the less popular products (i.e. those belonging to the long-tail) and recommend products that are similar to what a customer has already purchased and rated. The difference is in its traversal route. By using *Absorbing Time*, the *random walker* traverses through the unpopular nodes until it arrives at a node that represents the popular one. The detail of the recommendation process through *Absorbing Time* is presented in Algorithm 2.

#### IV. DATA AND EVALUATION

##### A. Data Set

This study adopts the approaches from previous studies [14], [17] by employing the tripartite graph representation to build the recommendation model for a B2C e-commerce

**Algorithm 2.** Recommendation by using Absorbing Time

**Input:**

A tripartite graph  $G = (V, E)$

A customer node  $q \in V$

Time  $t$  for how long the *random walker* traverse the nodes

**Recommendation\_By\_AT( $G, q, t$ ):**

- 1) define a subgraph  $G' = (V', E')$
- 2) define  $S \subseteq V'$  and  $S' \subseteq V'$  such that  $V' = S \cup S'$
- 3) for each product  $i$  that has been rated by customer  $q$ :
- 4) include node  $i$  as the member of  $S$
- 5) include edge  $(q, i) \in E$  as the member of  $E'$
- 6) for each product  $j$  that has not been rated by customer  $q$ :
- 7) include node  $j$  as the member of  $S'$
- 8) create a stochastic transition matrix  $M_S$  from  $G' = (S \cup S', E')$
- 9) perform dot products  $(M_S)^t$  represents the random walk length  $t$
- 10) for each node  $i$  in subgraph  $G'$ , calculate the *Absorbing Time* value:
- 11) if node  $i \in S$  then:
- 12)  $AT(S|i) = 0$
- 13) else:
- 14)  $AT(S|i) = 1 + \sum_{j=1}^n p_{i,j} \cdot AT(S|j)$
- 15) sort the *Absorbing Time* value in ascending order for all node in  $S'$

$$Recall@N = \frac{\sum hit@N}{|L|} \quad (6)$$

2) *Evaluation on Products Diversity*: The purpose of this measurement is to identify the recommendation performance in term of a variety of products, whether the recommendation covers both the popular and unpopular products or only focus on the popular ones. The higher the diversity value, the more different types of products would be recommended to customers.

The value of *Diversity* is obtained from the comparison of the number of unique products being recommended by the system and the maximum amount of the recommendation. The amount of the recommendation is calculated from the multiplication of the desired number of top recommendations with the number of customers involved in the experiment. The formula for *Diversity* score is presented in Equation (7).

$$Diversity = \frac{|\bigcup_{u \in U} R_u|}{|I|} \quad (7)$$

3) *Evaluation on Long-Tail Products*: The last evaluation utilizes the *Long Tail* measurement. This metric determines whether the recommendation system successfully recommends long-tail products. Rather than using the average rating as conducted in [2] which could lead to the misleading implication, this metric is modified by considering the average number of customers who give a rating of a product. For example, product A is purchased by five people and all of them give 1 rating. Product B is only purchased by one person and the given rating is 5. If the *Long Tail* score is calculated using the average value, then product A will be considered as a long-tail product while product B is the popular product. Equation 8 presents the equation for calculating the *Long Tail* score. The notation  $|L|$  denotes the number of tests, while  $rating(i)$  denotes the number of customers who give a rating on product- $i$ .

$$LongTail = \frac{\sum_{i=1}^n rating(i)}{|L|} \quad (8)$$

TABLE I. GRAPH DATA SET DESCRIPTION

| The Number of               | 1st-level Category | 3rd-level Category |
|-----------------------------|--------------------|--------------------|
| Total nodes                 | 63,068             | 3,011,244          |
| Category nodes              | 21                 | 626                |
| Product/item nodes          | 20,000             | 20,000             |
| Customer/user nodes         | 43,047             | 43,047             |
| user-item edges             | 80,000             | 80,000             |
| item-category edges         | 469,036            | 1,123,096          |
| user-category edges         | 1,808,148          | 1,808,148          |
| Average degree of each node | 37                 | 47                 |

company. One of the differences is that in this domain, the product category has several levels. Thus, there are two types of data set being used in this study, differentiated based on the level of the category as summarized in Table I. There are in total 63,068 nodes that are connected to the 1-st level category nodes and 3,011,244 nodes connected to the 3-rd level category nodes. The objective of this differentiation is to identify the effect of the specialization (by using the 3rd-level category) and generalization (by using the 1st-level category) in the latent information for the recommendation result.

**B. Evaluation Metrics**

This study uses three metrics to evaluate the performance of the recommendation system adopting from [2]. These metrics evolved the evaluation of the accuracy, diversity, and exposure of the long-tail products. For each of these criteria, the comparison of *Hitting Time* and *Absorbing Time* performance are evaluated.

1) *Evaluation on Accuracy*: This metric use *Recall@N* that measures the accuracy of the recommendation result for each algorithm (*Hitting Time* and *Absorbing Time*). It evaluates how far the algorithm could recommend the long-tail products.

Given a collection of products consisting of the combination of customers' favorite products and other randomly chosen products, the recommendation system would recommend top  $N$  recommendations. If a customers' favorite product is included in the top  $N$  recommendation, the value of *hit@N* would be 1 and 0 otherwise. The notation  $|L|$  represents the number of tests case, i.e. the total instances of long-tail products that are tested their membership to the top  $N$  recommendations. The formula for this metric is given in Equation (6).

**V. RESULT AND ANALYSIS**

Starting by entering a username of a customer into the system, a set of  $N$  products are generated as the recommended products for the customer, following Algorithm 1 (by using *Hitting Time*) and Algorithm 2 (by using *Absorbing Time*) separately. Each Algorithm is run by using two types of data set described in Subsection IV-A combined with a specific value of  $N$  (the number of products to be recommended) and  $t$  (the length of time needed by the *random walker*). The evaluation is conducted through the observation of the accuracy, diversity, and long-tail measurement.

**A. Evaluation on Accuracy**

The aim of this experiment is to identify the accuracy of the utilization of *Hitting Time* and *Absorbing Time* in different types of category levels, i.e the use of 1-st level category vs 3-rd level category. In general, the value of *Recall@N* is increasing with the increase of the number of products ( $N$ ),

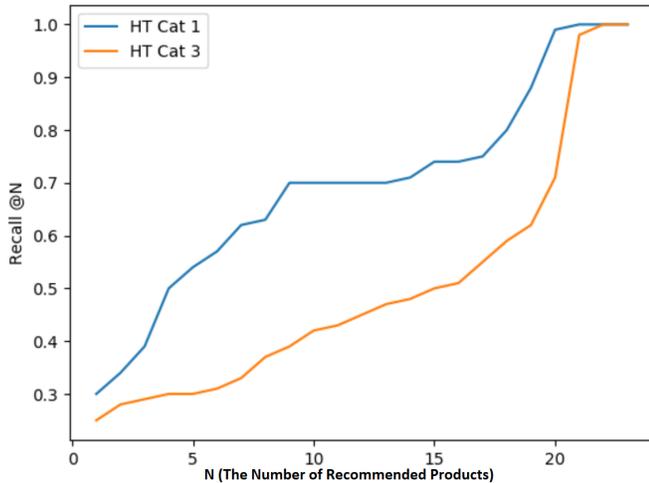


Fig. 2. The Accuracy (Recall@N) Value of using the Hitting Time.

but different combinations of the algorithm and category level being used lead to different results.

Fig. 2 present the accuracy comparison by using different level of category in the implementation of *Hitting Time*. This result shows that the accuracy of using the general category (1st-level category) is better than the specific one. This is caused by the condition that the 1st-level category has a smaller number of *user-category* edges. This situation produces a higher rating average on each of these edges that drives to the higher probability value in the transition matrix. Thus, the *random walker* traverses the graph faster to reach the nodes around the customer query node, especially when the nodes are in the same category.

Fig. 3 presents the comparison of accuracy in different category levels by using *Absorbing Time*. If the use of *Hitting Time* resulting better accuracy when it is combined with the use of data from the 1st-level category, *Absorbing Time* performs better in its combination with the data from the 3rd-level category. This difference is affected by the working principle of both algorithms in determining the source and target nodes to be traversed by the *random walk* and the different characteristics of the connectivity in each level of category.

Regardless of the level of category data, both results presented in Fig. 2 and 3 shows that the curve of the *Absorbing Time* is higher than the *Hitting Time*. This implies that the implementation of *Absorbing Time* has a better performance in terms of its accuracy. Since *Absorbing Time* would run better in a graph with a shorter path between nodes, this condition is consistent with the fact that the data set consisting more customer nodes than the product nodes. Therefore, the connectivity between product nodes has a shorter path.

### B. Evaluation on Diversity

In term of diversity, Fig. 4 shows the diversity comparison in top  $N$  recommendation ( $N = 5, 10, 15, 20$ ). This figure shows that the implementation of *Absorbing Time* tends to

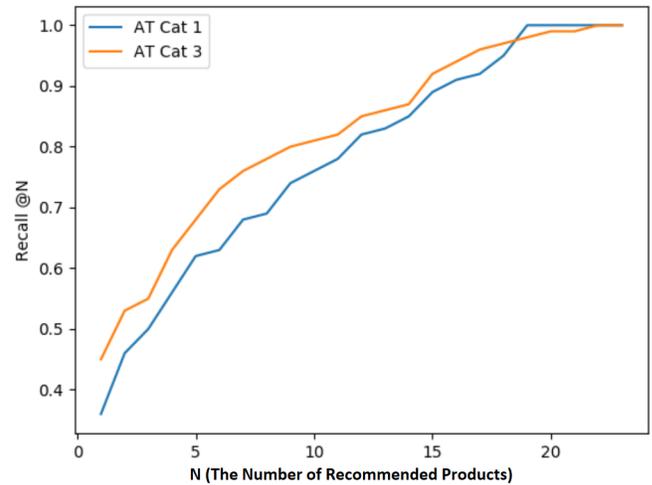


Fig. 3. The Accuracy (Recall@N) of using the Absorbing Time.

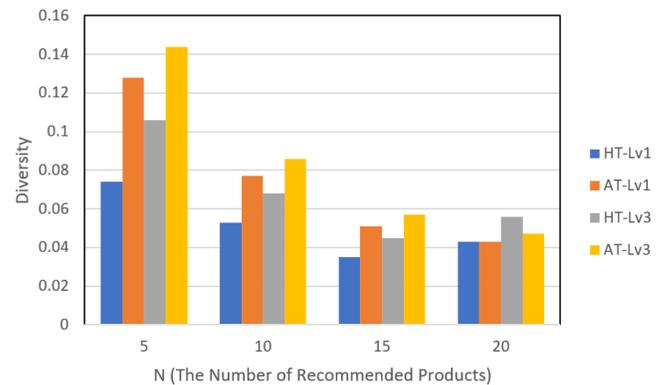


Fig. 4. The Comparison of Diversity Value by using *Hitting Time* and *Absorbing Time*.

be better than the *Hitting Time* in the small number of  $N$ . Moreover, the use of specific category data (i.e. 3rd-level) yields better diversity in both algorithms. From Fig. 4, it can be said that the greater value of  $N$ , the lower the diversity value would be obtained (for both approaches). However, this diversity value does not guarantee the quality (accuracy) of the recommendation because the accuracy is inversely proportional to diversity. Moreover, as summarized in [17] that the higher diversity value reflects the high probability of the long-tail products to appear in the recommendation. Thus, the *Hitting Time* algorithm produces the more diverse products, and probably captures the long-tails better than *Absorbing Time* algorithm, but might not be better in terms of accuracy.

TABLE II. THE LONG TAIL COMPARISON BY USING *Hitting Time* AND *Absorbing Time*

| Top $N$ Recommendation | <i>Hitting Time</i> | <i>Absorbing Time</i> |
|------------------------|---------------------|-----------------------|
| 5                      | 362.01              | 347.13                |
| 10                     | 366.18              | 333.19                |
| 15                     | 330.87              | 330.57                |
| 20                     | 327.01              | 322.01                |

### C. Evaluation on Long Tail

This evaluation calculates the average number of customers who give a rate to a product in each of the top recommendation levels. The lower average value denotes the better long-tail products recommendation. As presented in Table II, the value by using *Absorbing Time* implementation is slightly lower than the *Hitting Time*. This indicates that the more recommended products generated by the implementation of *Absorbing Time* come from the long-tail products. Compared to the result on diversity evaluation that *Hitting Time* probably has the more long-tail products to be recommended, from the result of the long-tail evaluation, it is not valid. Though the long-tail values between both algorithms are slightly different, the *Absorbing Time* performs better. Thus, as the aim of the long tail measurement is to identify whether the recommendation system correctly recommends more long-tail products, it is confirmed for the use of *Absorbing Time*.

## VI. CONCLUSION

This study focus to solve the long-tail problem specifically for B2C e-commerce domain using a tripartite graph representation. Markov random walker is employed to traverse the graph based on *Hitting Time* and *Absorbing Time* algorithm in order to recommend the products for the customers. The experimental result shows that *Absorbing Time* algorithm yields better accuracy than the *Hitting Time*. The use of this method also slightly generates more long-tail products to be recommended. In terms of diversity, the *Hitting Time* algorithm provides slightly more diverse recommended products. In addition, specialization and generalization on the product category levels as the latent information are observed. The experimental result shows that there is a difference in using generalized vs specialized category levels. *Absorbing Time* perform better in recommendation accuracy combined with the 3-rd level category, and in terms of diversity, the use of this specialized category level for both approaches shows the more diverse recommended products. This experiment shows that to deal with the problem of long-tail in the e-commerce domain, it is possible to make a recommendation by involving the products from the long-tail groups. The diversity score implies that the use of the more specific categories generates the more varied products to be recommended to the users. Through the implementation of the tripartite graph, either *Hitting Time* and *Absorbing Time* approach for graph traversal are considerably to be implemented in B2C companies.

## ACKNOWLEDGMENT

This work was supported by Universitas Indonesia through grant "Publikasi Terindeks Internasional (PUTI) Q2: NKB-4060/UN2.RST/HKP.05.00/2020"

## REFERENCES

- [1] V. A. Zeithaml, M. J. Bitner, and D. D. Gremler, *Services Marketing Strategy*. American Cancer Society, 2010.
- [2] H. Yin, B. Cui, J. Li, J. Yao, and C. Chen, "Challenging the long tail recommendation," *Proc. VLDB Endow.*, vol. 5, no. 9, p. 896–907, 2012.

- [3] C. Anderson, *The Long Tail Why The Future Of Business Is Selling Less Of More*. New York, NY: Hyperion, 2006.
- [4] A. Luke, J. Johnson, and Y.-K. Ng, "Recommending long-tail items using extended tripartite graphs," in *2018 IEEE International Conference on Big Knowledge (ICBK)*, 2018, pp. 123–130.
- [5] M. Aprilianti, R. Mahendra, and I. Budi, "Implementation of weighted parallel hybrid recommender systems for e-commerce in indonesia," in *Proceedings of the International Conference on Advanced Computer Science and Information Systems*. Malang - Indonesia: IEEE, 2016, pp. 321–326.
- [6] P. H. Aditya, I. Budi, and Q. Munajat, "A comparative analysis of memory-based and model-based collaborative filtering on the implementation of recommender system for e-commerce in indonesia: A case study pt x," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2016, pp. 303–308.
- [7] S. M. Rezaeinia and R. Rahmani, "Recommender system based on customer segmentation (rscs)," *IEEE Transactions on Knowledge and Data Engineering*, vol. 45, no. 6, pp. 946–961, 2016.
- [8] F. Rodrigues and B. Ferreira, "Product recommendation based on shared customer's behaviour," *Procedia Computer Science*, vol. 100, pp. 136–146, 2016.
- [9] R. Trivonanda, R. Mahendra, I. Budi, and R. A. Hidayat, "Sequential pattern mining for e-commerce recommender system," in *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2020, pp. 393–398.
- [10] J. Li, K. Lu, Z. Huang, and H. T. Shen, "On both cold-start and long-tail recommendation with social data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 1, pp. 194–208, 2021.
- [11] S. Wang, L. Hu, Y. Wang, X. He, Q. Z. Sheng, M. A. Orgun, L. Cao, F. Ricci, and P. S. Yu, "Graph learning based recommender systems: A review," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 2021, pp. 4644–4652.
- [12] A. A. Putra, R. Mahendra, I. Budi, and Q. Munajat, "Two-steps graph-based collaborative filtering using user and item similarities: Case study of e-commerce recommender systems," in *2017 International Conference on Data and Software Engineering (ICoDSE)*, 2017, pp. 1–6.
- [13] M. Jamali and M. Ester, "Trustwalker: A random walk model for combining trust-based and item-based recommendation," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 397–406.
- [14] X. Hu, C. Zhang, M. Wu, and Y. Zeng, "Research on long tail recommendation algorithm," in *Proceedings of the International Conference on Artificial Intelligence Applications and Technologies*, vol. 261. IOP Publishing, oct 2017, pp. 012–019.
- [15] Y. Wang, J. Wang, and L. Li, "Enhancing long tail recommendation based on user's experience evolution," in *2018 IEEE 22nd International Conference on Computer Supported Cooperative Work in Design ((CSCWD))*, 2018, pp. 25–30.
- [16] Y. Wang, J. Wu, Z. Wu, H. Yuan, and X. Zhang, "Popular items or niche items: Flexible recommendation using cosine patterns," in *2014 IEEE International Conference on Data Mining Workshop*, 2014, pp. 205–212.
- [17] J. Johnson and Y.-K. Ng, "Using tripartite graphs to make long tail recommendations," in *2017 8th International Conference on Information, Intelligence, Systems Applications (IISA)*, oct 2017, pp. 1–6.
- [18] A. Iványi, S. Pirzada, and F. A. Dar, "Tripartite graphs with given degree set," *Acta Universitatis Sapientiae, Informatica*, vol. 7, no. 1, pp. 72–106, 2015. [Online]. Available: <https://doi.org/10.1515/ausi-2015-0013>
- [19] L. László, "Random walks on graphs: A survey, combinatorics, paul erdos is eighty," *Bolyai Soc. Math. Stud.*, vol. 2, 01 1993.
- [20] D. Aldous and J. A. Fill, *Reversible Markov Chains and Random Walks on Graphs*, 2002. [Online]. Available: <http://www.stat.berkeley.edu/~aldous/RWG/book.html>

# Machine Learning Applied to Prevention and Mental Health Care in Peru

Edwin Kcomt Ponce, Melissa Flores Cruz, Laberiano Andrade-Arenas  
Facultad de Ciencias e Ingeniería  
Universidad de Ciencias y Humanidades  
Lima, Perú

**Abstract**—The present research aims to develop an application that allows the early and timely detection of signs of problems in the mental health of citizens. Agile methodology was used, with its SCRUM framework developing its four steps. In addition, technological tools such as artificial intelligence, mobile applications, social networks and the python programming language were used. Also using SQL Server, Android Studio and the Marvel applications, the latter for the design of the prototypes, through the method of sentiment analysis and machine learning, in order to create a mobile application that is as accurate as possible in its results. For this, several types of algorithm were evaluated, managing to select the most appropriate one since it works based on information collected through the social networks Facebook and Twitter. The result that was obtained was the application that uses machine learning to prevent and take care of mental health in Peru, thus benefiting the citizens of society.

**Keywords**—Artificial intelligence; machine learning; mental health; scrum; sentiment analysis

## I. INTRODUCTION

The current global health emergency caused by the SARS-CoV-2 virus has left behind countless deaths and confirmed cases due to the high level of contagion of the disease. Since then, the countries have focused their efforts mainly on controlling and taking the necessary preventive measures to avoid an increase in the number of infections by implementing public health policies [1]. At the same time, various academic articles regarding the coronavirus have been disseminated, the vast majority focusing on epidemiological, genetic and clinical aspects. However, know that the pandemic caused by COVID-19 [2], also brought uneasiness and negative thoughts in the general population that have been spreading since then threatening mental health.

Peru has been one of the countries most affected by the pandemic. Thus, the population had to respect the restrictions established by the governments of the day, remaining long periods of confinement, which in the short or long term represents a potential psychosocial impact on children, adolescents and adults. Due to a radical change in their way of life and the stressful environments caused by the pandemic, as a consequence, concern about the consequences arises not only on physical and mental health [3]. In a survey conducted during the current juncture of 546 people [4], the result was that 69.2% of the participants showed changes in behavior and emotions as well as high levels of stress (47.2%).

In addition to all the above and before the start of the pandemic, in Peru it is estimated that 1 in 5 individuals shows signs of a significant mental problem each year. In the last

4 decades, various initiatives have been promoted but these have not been sustainable, it is not until the approval of Law 29889 that a series of innovations in the provision of services take place in the period 2013-2018. All this has made it possible to implement more than a hundred new mental health care facilities [5]. However, everything is still subject to the commitment of the authorities, the gradual increase in public financing and joint national and international strategies.

In addition, a study of burden by disease category carried out in the country already revealed that mental and behavioral disorders were at the top of the list. It is estimated that approximately 20% of the adult and older population suffer from mental health problems, especially depression, anxiety and alcohol dependence. In addition to the above, 20% of boys and girls suffer from behavioral and emotional disorders; these antecedents are the consequence of an insufficient supply of mental health services by the state. The resources are still insufficient to manage to address the previous problems such as the new ones caused by psychosocial stress [6], linked to exposure to a possible contagion, insecurity and prolonged confinement.

Faced with this pandemic and the various challenges it poses in health matters, the College of Psychologists of Peru consider that in the face of this new reality it is important to be able to adapt to the changes and challenges country is going through. In this sense, they state that the use of technologies represents an opportunity to update knowledge with new learning and various challenges that promote research [7]. Therefore, in order to face mental health problems, it is important to promote mental health research because the country is at risk of continuing to increase the burden of disease from psychiatric institutions. So the need arises for this area to be included in the priorities of health research in Peru [8], the latter being recently ratified by the Ministry of Health for the period 2019-2023.

It is for these reasons that it is important to offer support to contribute and improve the efficiency of mental health plans promoted by the competent institutions. Thus, achieving to offer innovative ways of attending to mental health in the country through the use of technology, thus reinforcing the still fragile efforts to provide a service in conditions to the population. In this way, the effectiveness of the care of our professionals was increased, lightening the load of the functions they perform for better performance.

The objective of this work is to implement an app for the prevention and care of mental health in the country through the

use of artificial intelligence. This app can be downloaded to any Smartphone which gives greater accessibility to the service since nowadays everyone has a mobile device. The purpose is to quickly anticipate possible reactions caused by mental illnesses and disorders, the application by means of the activity record in messaging services and / or social networks allow knowing the current situation of the user. If any risky behavior is detected, an alert notification was sent to 1 or more trusted persons. In addition, it was possible to know the mood on a daily basis, which is very important for a person with mental health problems, the latter was possible through a rating scale that have to be marked at the beginning of the day, which allows better monitoring and follow-up user.

## II. LITERATURE REVIEW

In carrying out the research work, the subject of artificial intelligence was addressed and how this technology applied to health generates significant contributions to improve the precision and effectiveness of diagnoses and treatments of diseases. It focus mainly on mental health in the country, becoming a very helpful support for the work of specialized medical personnel.

According to the author [9], comments that suicide as a mental health problem is increasing worldwide with a figure of approx. 800,000 deaths per year added to a subjective suicide risk assessment process which limits its efficacy and accuracy. Due to this, suicide detection strategies are focusing on the use of artificial intelligence for the optimization of suicide risk prediction and behavior management. The methodology use was based on articles published between the years 1990 to 2019 and how artificial intelligence has had a positive impact on the care and monitoring of mental health. In addition, artificial intelligence has been used to support the clinical management of suicide, demonstrating the advantages of incorporating this technology. Ideal for use in remote locations with limited access to mental health care. The author concludes based on the observed benefits artificial intelligence has a proven advantage for suicide prediction and mental health care.

On the other hand the author [10], confirming that existing mental disorders have been even more affected by the current global situation caused by COVID-19 increasing mental health problems. To better understand the role of artificial intelligence in their research, the methodology used was based on a review of 253 articles. In the analysis of his framework, he consisted in deriving ideas, concepts and knowledge that was integrated into the development of his project. The results obtained were mainly the possible applications of telepsychiatry and artificial intelligence as well as characteristics and models of artificial intelligence in mental health. Finally, they conclude that even these new technologies cannot be fully adopted in the field of mental health care, they consider that health professionals must choose the most appropriate tools. Based on various aspects, a balance must be found between conventional care and technology-based care, which was achieved progressively.

It also coincides that in order to assess and treat the sequelae of mental health and possible psychiatric comorbidities, it is important to optimize patient care [11]. To ensure the efficient use of limited resources, artificial intelligence can help to achieve this. It considers for its methodology the use of

artificial intelligence applications that include validations based on clinical trials. As a result, it is evident that most artificial intelligence applications use simulated data sets that limit the rate and restrict its applicability in a clinical population and in a real world environment. In conclusion, however, more up-to-date and innovative test designs can generate better data sets that are generalizable to the entire world population. The acquisition of large volumes of data is of the utmost importance as they are vital to guarantee that the applications allow to obtain greater precision in the results, especially when they are used as part of a diagnosis or clinical treatment.

The objective of the study was a preliminary evaluation of real-world data to verify the effectiveness of a mental well-being mobile application that interacts through text with users with symptoms of depression [12]. The methodology put into use for this case was to observe a group of anonymous users who installed the application. The study used a mixed methods approach to evaluate the impact and levels of user participation. Quantitative analysis measured the impact of the app by comparing the average improvement in depression symptoms among users. The results obtained were the average improvement in the participants' state of mind, and 67.7% considered the application experience useful and encouraging. The effectiveness and levels of user participation were concluded as promising; however, these first findings have yet to be validated in much larger samples and over longer periods of time.

In addition to the aforementioned, he agrees that the current situation has aggravated various existing problems in the field of health and has focused on knowing a list of factors that could show a predisposition to a mental disorder [13]. For its methodology, a survey was conducted of 17,764 adults of different age groups, genders and economic status through statistical analysis and Bayesian network inference. Key factors that affected the mental health of the participants were identified during the pandemic, the integration of Bayesian networks with classic machine learning approaches allowed generating an effective model of the prevalence level of mental health. It was also recognized who were more prone to mental disorders and causes that cause mental pressure, with the aforementioned it was prefixed with a precision of approx. 80% that people are more mentally vulnerable. As a conclusion, it was determined that factors influence mental health problems during the pandemic and what activities help to keep at bay from disorders that may affect, confirming that caring for people with a history of mental illness seems to be more important during this time.

Finally, after what has been stated by the various authors mentioned, consider that the use of artificial intelligence is of utmost importance and is relevant for a faster, more efficient and optimal work with regard to mental health as it shows great advantages that serve as support health professionals in their work, reducing the workload which not only help better decision-making, but also ensures better care and more accurate results in diagnoses and treatments as well as monitoring and supervision of the patient until recovery.

## III. METHODOLOGY

Next proceed to describe the methods and tools used for this research work.

### A. SCRUM Methodology

This framework implies a process management to tackle complex projects that require dynamic environments, so they demand speed of results and flexibility. It is an agile work methodology that has several objectives and that allows to accelerate processes, act quickly on possible changes and make periodic deliveries of work.

1) *Beginning*: In this first phase, the respective roles are identified and assigned, who was the Scrum Master and the Stakeholders [14]; the members of the work teams are defined according to their abilities and contributions to the project.

2) *Planning and Estimation*: In this second phase, the user stories are created, the sprint backlog [15], also be carried out taking into account the estimate for their correct implementation of each one of them [16].

3) *Implementation*: In this third phase, the respective prototypes are designed, taking into account the requirements of each of the stories raised.

4) *Reviews and Retrospectives*: In this fourth phase, the respective review of the Sprint is carried out by the team, in an activity that allows the inspection and adaptation of the product, the most important thing is the conversation by the team to understand the situation and receive feedback [17]. In Fig. 1 observe the order of the processes for the Scrum methodology.

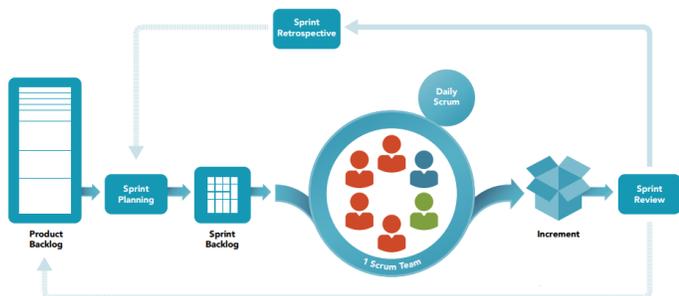


Fig. 1. Scrum Methodology Flow.

### B. Technological Tools

The technologies used for this research work guarantee an optimal development of the mobile application based on the knowledge of the members, managing to finish it according to the specifications and in the established time.

1) *Artificial Intelligence*: This technology seeks to provide software with the capacity for learning based on data, with the latter patterns and opportunities arise through which performance tests are carried out to measure efficiency based on the percentage of errors and successes. The present research project opts for machine learning as it is not linear since it adapts through learning to new cases or situations, even not needing constant supervision. In Fig. 2 can see the steps of machine learning.

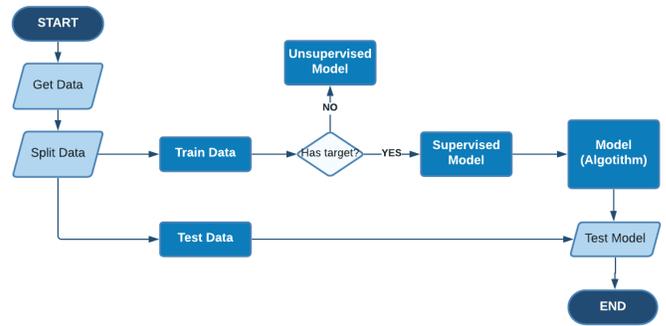


Fig. 2. Flow Diagram of the Machine Learning Modeling Process.

2) *Mobile Apps*: Refer to programs designed for exclusive use on mobile devices, generally smartphones, which allows the user access to various services and functions, both professional, entertainment, educational, among others, facilitating the development of their activities.

3) *Social Networks*: Social networks are communities made up of various people and organizations that relate to each other through internet access, allowing immediate communication between users, sharing information and leisure activities. For the present work, take advantage of the messaging service that these communities provide, serving as the basis for the analysis of the individual's behavior and considering the massive use of these tools. Table I shows the results of a survey carried out by IPSOS [18], confirming the high percentage of use of social networks in the country.

TABLE I. USE OF SOCIAL NETWORKS IN PERU 2020

| Estimated Users | Population | Most used social networks                                   |
|-----------------|------------|-------------------------------------------------------------|
| 13.2 Millions   | 78%        | Facebook, WhatsApp, YouTube, Instagram, Messenger, Twitter. |

4) *Python*: It is a high-level programming language that manages to process all kinds of data structures, whether text or numeric. It has taken the characteristics of its predecessors, it is free software, that is to say, open source and allows it to be used and distributed freely even for commercial use. It is accessible, simple and multiplatform, for this project it represents a great advantage due to its wide library, selection of frameworks and its simplicity in syntax.

### C. System Requirements

The programs necessary for an adequate development environment are considered for this project, allowing the realization and implementation of the app correctly.

1) *SQL Server*: is a relational database management system (RDBMS). It supports a wide variety of transaction processing, business intelligence, and analytical applications in corporate IT environments.

2) *Android Studio*: is a development platform that allows you to build mobile applications exclusively for Android operating systems, as well as various tools that allow you to develop an app that is stable on the target devices.

3) *Marvel App*: is an online tool to make layout and prototypes of both web pages and applications on mobile devices.

#### IV. CASE STUDY

##### A. Planning

This project is made up of several stages where various advances are made that must be completed in a certain time. These stages begin with the definition of user stories that allow identifying the system requirements. The project has 10 user stories with an approximate time of 15 weeks for the stories to be finalized. In Table II can see the description of each story, which allows to better understand the functions that the app has.

TABLE II. USER HISTORY

| History No. | Description                                                                                                                                                                                                                    |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| H1          | As an administrator I want the application to show a friendly and intuitive interface so that it is easy to use and allows the user to register.                                                                               |
| H2          | As an administrator, I want the application to ask the user to mark their state of mind on a daily basis.                                                                                                                      |
| H3          | As an administrator, I want the application to be able to collect information from social networks and messaging services used by the user.                                                                                    |
| H4          | As administrator I want the application to send a daily report to the email of the assigned medical specialist.                                                                                                                |
| H5          | As an administrator, I want the app to display a questionnaire on the cell phone screen that the user must fill out on a mandatory basis in order to better control their mental health.                                       |
| H6          | As a user, I want to be able to register the contact details of trusted people.                                                                                                                                                |
| H7          | As a user, I want the application to show me information related to emotional well-being when entering the app.                                                                                                                |
| H8          | As an administrator, I want the application to send an alert notification to the user's trusted persons in case risky behavior is detected.                                                                                    |
| H9          | As an administrator, I want the application interface design to be made based on colors and sounds that transmit calm and stimulate positive emotions in the user.                                                             |
| H10         | As an administrator, I want the application to show the user motivational and / or informative messages as a preventive function in case indications of possible risky behaviors are detected without having to enter the app. |

##### B. Estimate

For this phase, the user stories that have been previously defined are organized through the product backlog allowing to have fixed goals and meet the established deadlines. The estimations are made by means of the planning poker technique where the team is in charge of assigning a number (Fibonacci series) to each user story to be able to classify them through previous agreements. As can be seen in Table III in the estimation column, the team determines the effort involved in developing each function and requirement raised based on history 6 since its development is the one with the least difficulty.

To define the priority of user stories, take into account the effort and difficulty that their respective development, operation and execution may require, as well as the importance and relevance that it supposes for the present project.

The development of the project is divided into three sprints, with the first sprint obtaining a total of 8 story points, being

TABLE III. PRODUCT BACKLOG

| History No. | Estimate | PRIORITY | Sprint |
|-------------|----------|----------|--------|
| H3          | 5        | 1        | 3      |
| H4          | 5        | 2        | 3      |
| H10         | 8        | 3        | 3      |
| H8          | 8        | 4        | 2      |
| H5          | 3        | 5        | 2      |
| H6          | 1        | 6        | 2      |
| H2          | 2        | 7        | 2      |
| H1          | 2        | 8        | 1      |
| H7          | 3        | 9        | 1      |
| H9          | 3        | 10       | 1      |

the one that requires the lowest speed, which allows the team to integrate and gain greater confidence in the process and progress of the sprint in question. The second and third sprints receive 14 and 18 points respectively and although the points have increased, the team is already able to organize and communicate much better, which allows a greater understanding and minimizes errors when working as a team. In Fig. 3, see in a graph the sprints already organized and with their respective story points.

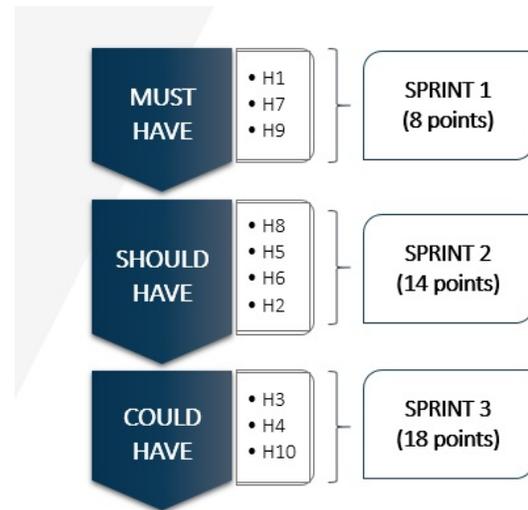


Fig. 3. Story Points and Number of Sprints.

##### C. Implementation and Development

In this phase, the process to develop the mobile application is detailed using the technologies and procedures proposed. The modeling of the Machine Learning software is carried out, in addition [19], at this stage make the choice of the algorithm for learning the artificial intelligence software.

1) *Sentiment Analysis*: It is a process that analyzes opinions, behavior and impressions of users, basically it consists of extracting valuable information after having evaluated attitudes and emotions behind a series of words, focusing on the lexicon that expresses feelings. It has a generally commercial use with applications in marketing, politics, services, companies, surveys, brand positioning, etc. For the present project [20], this method is used for the analysis of data in social networks because the latter have gone from being simple means of

communication to means of critical thoughts and / or opinion, allowing the content to be classified into two categories: positive and negative.

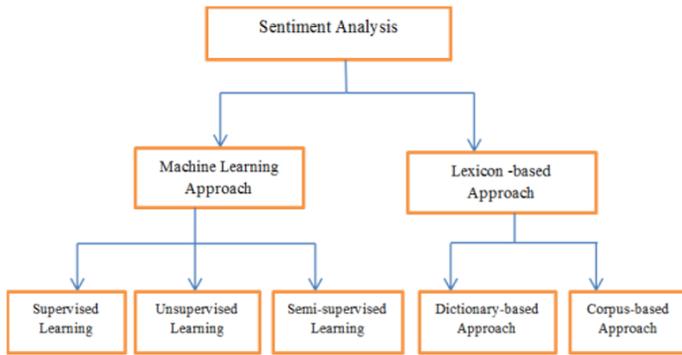


Fig. 4. Methods for Sentiment Analysis [21].

In Fig. 4, observe the two methods that can be used for the process and development of sentiment analysis, for the present work the alternative of an automatic learning model is chosen to achieve better precision and accuracy of the results.

2) *Supervised Learning*: Machine learning is divided into 2 types of methods for the application of the project, supervised learning is chosen since it is the most recommended for jobs that require the use of classification and [22], allowing developers to more accurately identify the processes of intelligence software artificial and at the same time have a better control of the training material for the learning process.

3) *Datasets*: The data set is divided into two parts, the first data is used for tests and the rest as training, the latter being used to achieve Machine Learning modeling that gives the most exact results possible and with the least margin of error.

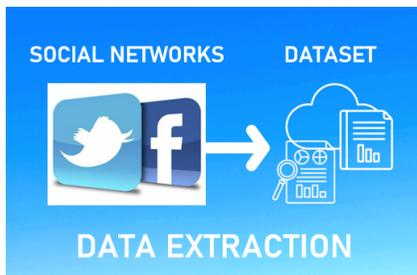


Fig. 5. The Data Extracted, Publications and Content Shared in a Public Way.

In Fig. 5, Facebook and Twitter are considered as social networks of reference for the extraction of data from the present project. To obtain the necessary data, the API that the platforms offer for developers is used. The results obtained from the model [23], are compared with information based on the medical history of patients who have presented disorders related to depression and anxiety.

4) *Data Preprocessing*: The set of training and testing data for the algorithm is previously subjected to a series of techniques and procedures that allow cleaning and reducing

certain characteristics of the texts that are irrelevant for processing. The purpose of this phase is to normalize the data by converting the text into vectors for the classification process and facilitating sentiment analysis using Machine Learning.

a) *Filter*:

In some cases the data may contain special characters that denote admiration, questions or some reference to web pages or tagging. These characters must be removed obtaining a clean data set for vector representation and model classification.

b) *Tokenization*:

It allows to divide the sentences into smaller parts called tokens, which are used for later stages of the processing, facilitating the use of the data. For the project, a token is equivalent to a word.

c) *Anonymity of Personal Information*:

The identity of the authors of social media posts must be kept anonymous so as not to expose users without their consent.

d) *Remove Stop-Words*:

In natural language there are many words that are used frequently and that by themselves do not keep any meaning. In the Spanish language these words are usually articles, conjunctions and pronouns.

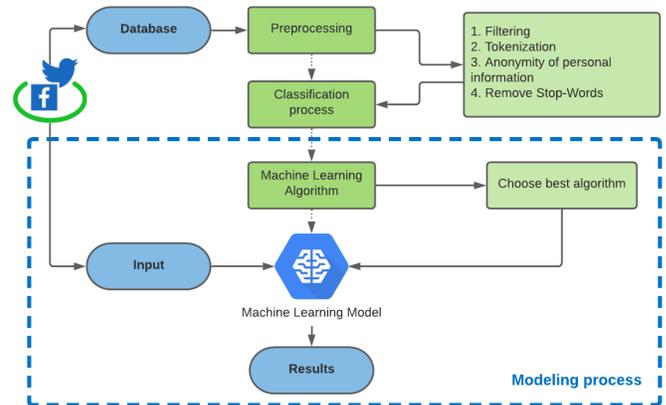


Fig. 6. Architecture for the Dataset Collection Process for the Testing and Training of the Machine Learning Model.

5) *Algorithm*: For this project, the Support Vector Machines classification algorithm is chosen, due to its advantage and ability to work with large databases in addition to performing text classification very well because it can handle large functions and in turn demonstrate robustness when the set of data is small and is distributed in a large area, for those reasons it has given reliable results in past research.

In Fig. 6, the process from when the data are obtained to their respective classification and modeling with the chosen algorithm is graphically represented. As mentioned in [24] correct data processing can make the difference between a model with lower or higher performance. This project takes these procedures into consideration to guarantee a model that is as accurate and precise as possible.

In Fig. 7, see a graph that represents the operation of the algorithm, dividing classes, that is, cases versus non-cases, based on a line called the hyperplane. The hyperplane

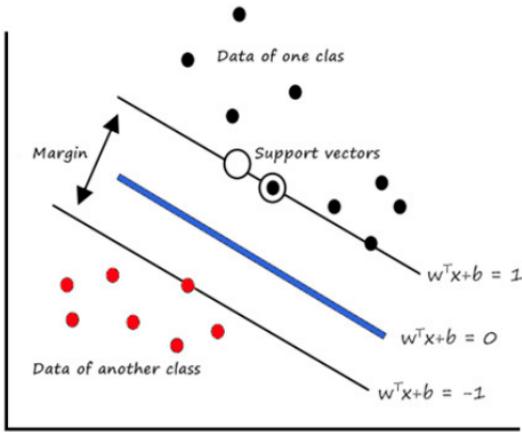


Fig. 7. Graphic Representation of the Algorithm Used Support Vector Machines [25].



Fig. 8. First and Second Prototype.

is created based on the greatest possible distance from the closest neighboring predictor data points between the classes [25]. More complex data that cannot be separated into two dimensions can be raised to a higher dimension through a process called kernelling.

## V. RESULT AND DISCUSSION

### A. Design and Prototypes

The designs are evidenced based on the sprints and user stories, each one is described to understand in greater detail the functions that are implemented in the mobile platform.

1) *First Sprint:* This sprint focuses mainly on the user interface and menu section, in Fig. 8 can see on the left the prototype that shows the options to be able to register and log in to the application, the use of sober colors is considered in the realization of the prototypes so that they can transmit calm stimulating positive emotions to the user.

In the prototype on the right note the data that the user is required to register and create an account, in turn it has a section where you fill in the data of trusted people or close contacts.

2) *Second Sprint:* This sprint focuses on the part of mental health care of the user, in Fig. 9 the alert message sent to the trusted person or close contact of the user is shown, in this way preventive measures are reinforced and maintained an alert to any risky behavior that can be detected by the app based on the data obtained from the activity on the smartphone.

In the following figure, a form appears on the user's screen that must be filled out monthly in order to have greater control over the behaviors, moods or behaviors that the user may manifest, said screen appear without the need for the user. Once entered into the app, the questions allow to detect symptoms related to depression or anxiety. In addition, an interface appears to the user at the beginning of his day, he must fill it with a mandatory character to know her state of mind, this screen appear daily.



Fig. 9. Third and Fourth Prototype.

3) *Third Sprint:* This sprint is related to the collection of information based on the user's interactions with their social networks and messaging services. In Fig. 10 observe the report that it generates monthly to be sent to the health specialist or medical center where the information was evaluated and allows the user's diagnosis and treatment to be further adapted.

### B. Mobile App

Once the model has been completed and put to the test, the application must now obtain the permissions to access the mobile device and begin to collect data on the user's activities through the use of their social networks and messaging applications. In Fig. 11, see graphically the steps that the application follows once it is installed on the mobile device and launched.

The information already being stored is examined by the artificial intelligence program and is responsible for comparing patterns through a database that determine indications or possible changes in the user's behavior and that may pose a

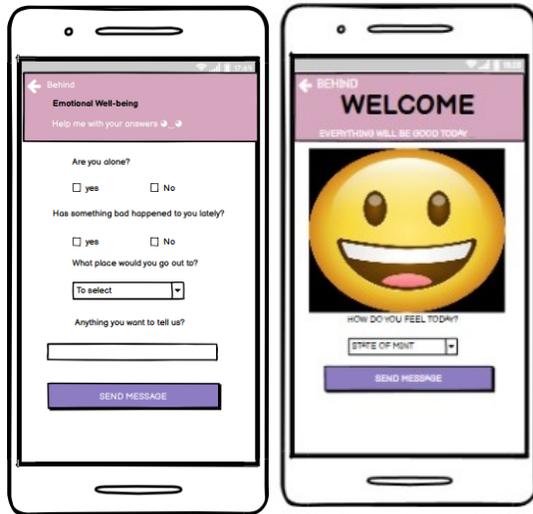


Fig. 10. Fifth and Sixth Prototype.

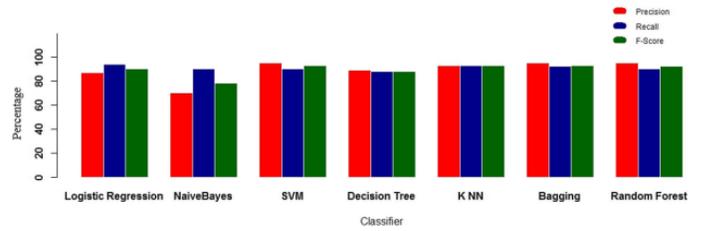


Fig. 13. Bar Graph Comparing the Performance of the Algorithm Proposed in [26].

In Fig. 13, the Super Vector Machines (SVM) algorithm stands out for the results obtained in precision and F-score (accuracy) obtaining 0.95% and 0.92% respectively, demonstrating good performance to predict users that may present problems of mental health. In addition, the good result of the algorithm is recognized in the face of this type of problem.

#### D. Mental Health in Peru

Mental health problems in the country have increased with the current situation, according to data from INFOSALUD telephone line authorized by the state for citizens who want psychosocial support, it is confirmed that the first 2 places of the most made calls are related to mental health problems where anxiety and depression occupy second place only in the months of April to May 2020 with a total of 3144 calls [27].

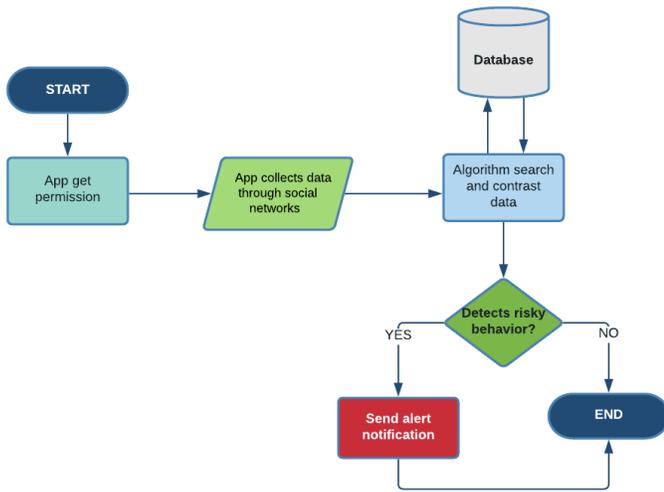


Fig. 11. Flow Chart of the Mobile Application.

risk to this or their environment. In case a risky behavior is detected, an alert notification is sent to trusted people who have been registered by the user in the application, being relatives or the assigned health specialist. In Fig. 12, observe the architecture of the app and the processes it carries out communicating through internet access to notify in case the algorithm detects unusual behavior on the part of the user.

#### C. Algorithm Performance

The present work proposes an application for the care and prevention of mental health, mainly focused on depression and anxiety disorders that have worsened even more with the current pandemic. On [26], the good performance of the algorithm chosen for this project is verified, where 3 aspects are taken into consideration: precision, sensitivity and F-score.

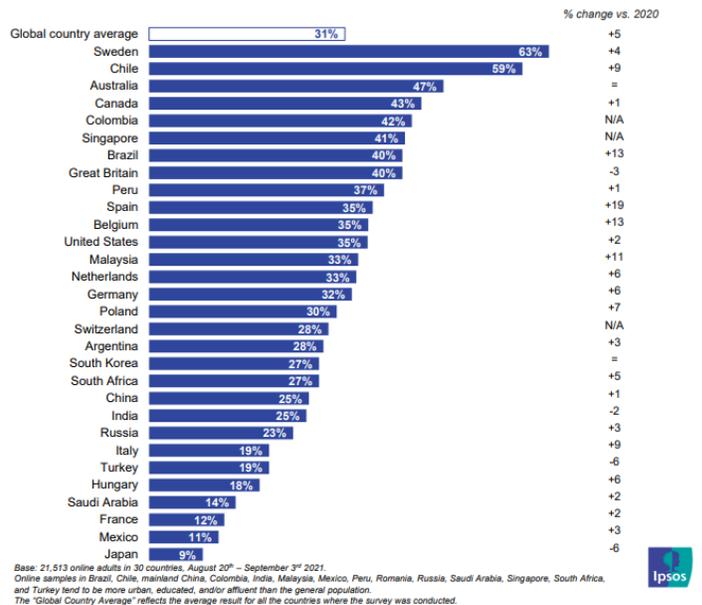


Fig. 14. Ipsos Survey 2020 [28].

In Fig. 14, a survey recently carried out by the IPSOS pollster is shown where 37% of those interviewed consider that one of the biggest problems facing the population is mental health.

The survey was carried out during the pandemic period, so the responses of the interviewees reflect the reality of a large part of the population that not only considers to have been affected by the economic situation but also psychologically

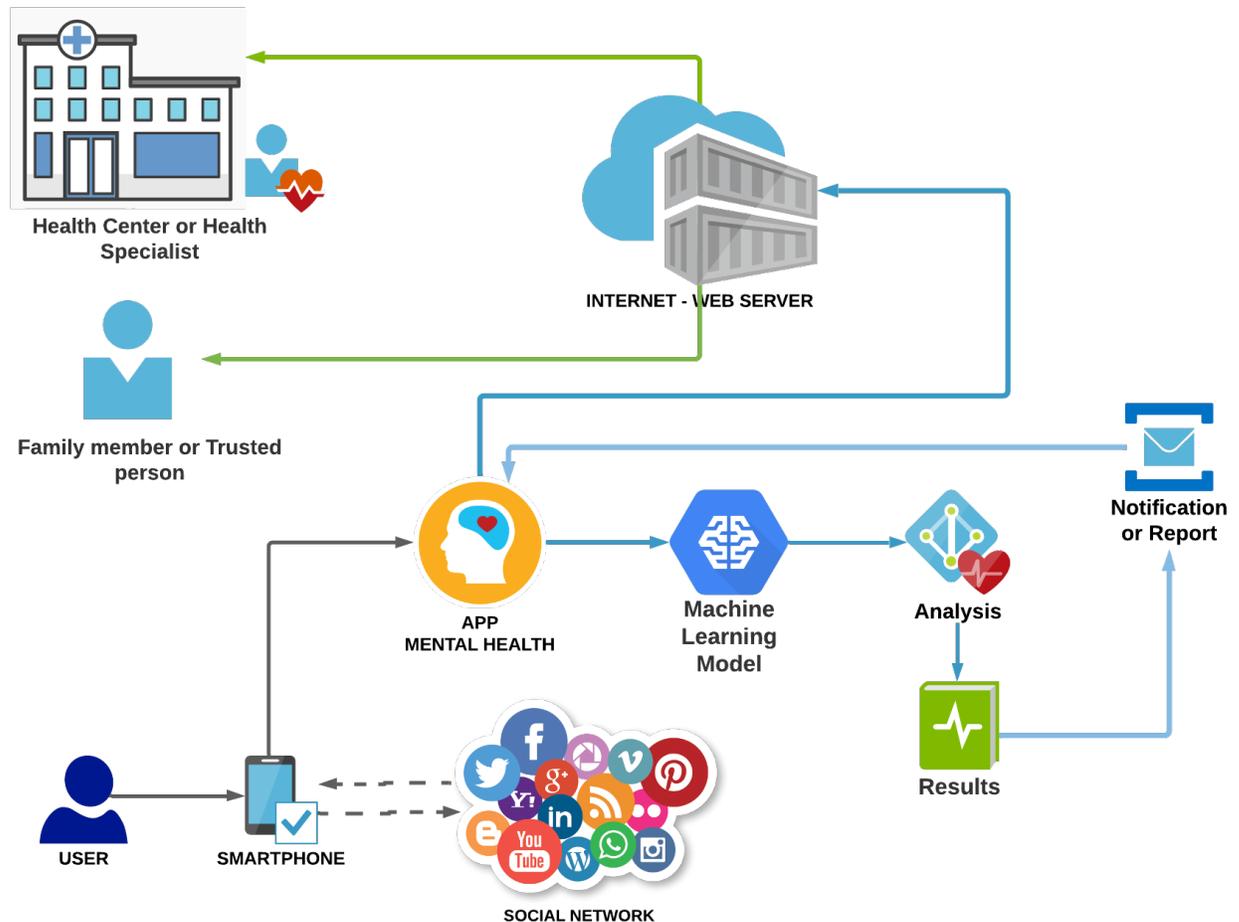


Fig. 12. Mobile Application Architecture.

by the long period of quarantine that still exists lives in the country, added to the fear and concern of being infected with the Covid-19 virus.

The results of this project favor to reduce the psychological effects caused by quarantine for this reason, the development of an application that serves as support to specialized health personnel turns out to be a support that generates a positive impact for both doctors and patients. In this way, access to personalized care is facilitated and appropriate preventive measures can be taken in the event of warning signs that may manifest as risk behaviors that are induced by mental health problems.

## VI. CONCLUSION AND FUTURE WORK

The research concludes that the application of Machine Learning allowed effective and efficient monitoring of mental health care and prevention through better control of user behavior through the use of social networks, even allowing the appropriate measures to be taken. for behaviors that indicate possible harm to your mental health. In addition, the scrum methodology allowed making suitable prototypes to be able to prevent health and thus benefit citizens. For future work, it is

recommended to continue increasing the data for the training of the Machine Learning model, allowing it to be more robust with information related to more mental health diseases and expanding its predictive capacity. This project has only been limited to exploring results considering an ideal environment for data collection, it is also recommended to improve the processing in cases where the data may not be objective, such as cases of sarcasm, spelling errors and spam in the content written by the users.

## ACKNOWLEDGMENT

The research that was carried out was supported by the University of Sciences and Humanities and its research institute.

## REFERENCES

- [1] J. Huarcaya-Victoria, "Mental Health consideration about the Covid-19 pandemic," *Revista Peruana de Medicina Experimental y Salud Pública*, vol. 37, no. 2, pp. 327–34, 2020.
- [2] J. Amachi-Choqqe and M. Cabanillas-Carbonell, "Iot system for vital signs monitoring in suspicious cases of covid-19," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120223>

- [3] M. Cabanillas-Carbonell, R. Verdecia-Peña, J. L. H. Salazar, E. Medina-Rafaile, and O. Casazola-Cruz, "Data mining to determine behavioral patterns in respiratory disease in pediatric patients," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120749>
- [4] F. Rusca-Jordán, C. Cortez-Vergara, B. C. Tirado-Hurtado, and M. Strobbe-Barbat, "Una aproximación a la salud mental de los niños, adolescentes y cuidadores en el contexto de la COVID-19 en el Perú," *Acta Medica Peruana*, vol. 37, no. 4, pp. 556–558, 2020.
- [5] E. N. E. L. C. Covid, E. Rita, M. Uribe, M. Psiq, V. Herrera, P. Gladys, and Z. Champi, "Plan De Salud Mental," vol. 2021, pp. 2020–2021, 2021.
- [6] Y. Castillo-Martel, Humberto; Cutipé-Cárdenas, "Simposio MENTAL HEALTH SERVICES REFORM IN PERU, 2013-2018," vol. 36, no. 2, pp. 2013–2018, 2019.
- [7] Colegio de Psicólogos del Perú, "El Colegio De Psicólogos Del Perú Y La Tele Psicología En Tiempos De Pandemia," no. 29733, pp. 2018–2020, 2021.
- [8] R. Valle, M. T. Rivera-Encinas, and S. Stucchi-Portocarrero, "Producción, impacto y colaboración en investigaciones peruanas en psiquiatría y salud mental," *Acta Medica Peruana*, vol. 37, no. 3, pp. 285–293, 2020.
- [9] T. M and A. Annamalai, "Telepsychiatry and the Role of Artificial Intelligence in Mental Health in Post-COVID-19 India: A Scoping Review on Opportunities," *Indian Journal of Psychological Medicine*, vol. 42, no. 5, pp. 428–434, 2020.
- [10] T. M. Fonseka, V. Bhat, and S. H. Kennedy, "The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors," *Australian and New Zealand Journal of Psychiatry*, vol. 53, no. 10, pp. 954–964, 2019.
- [11] G. Delanerolle, X. Yang, S. Shetty, V. Raymont, A. Shetty, P. Phiri, D. K. Hapangama, N. Tempest, K. Majumder, and J. Q. Shi, "Artificial intelligence: A rapid case for advancement in the personalization of Gynaecology/Obstetric and Mental Health care," *Women's Health*, vol. 17, 2021.
- [12] B. Inkster, S. Sarada, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, pp. 1–14, 2018.
- [13] I. P. Jha, R. Awasthi, A. Kumar, V. Kumar, and T. Sethi, "Learning the mental health impact of COVID-19 in the United States with explainable artificial intelligence: Observational study," *JMIR Mental Health*, vol. 8, no. 4, pp. 1–11, 2021.
- [14] V. Gomero-Fanny, A. R. Bengy, and L. Andrade-Arenas, "Prototype of web system for organizations dedicated to e-commerce under the scrum methodology," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120152>
- [15] R. Arias-Marreros, K. Nalvarte-Dionisio, and L. Andrade-Arenas, "Design of a mobile application for the learning of people with down syndrome through interactive games," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2020.0111187>
- [16] A. Ramos-Romero, B. Garcia-Yataco, and L. Andrade-Arenas, "Mobile application design with iot for environmental pollution awareness," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120165>
- [17] A. Tupia-Astoray and L. Andrade-Arenas, "Implementation of an e-commerce system for the automation and improvement of commercial management at a business level," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 1, 2021. [Online]. Available: <http://dx.doi.org/10.14569/IJACSA.2021.0120177>
- [18] Ipsos, "Uso de Redes Sociales entre peruanos conectados 2020 ," *Innovación y Conocimiento: Investigación de medios y comunicación de marca*, vol. 2020, p. 2020, 2020. [Online]. Available: <https://www.ipsos.com/es-pe/uso-de-redes-sociales-entre-peruanos-conectados-2020>
- [19] A. Biradar and S. G. Totad, *Detecting Depression in Social Media Posts Using Machine Learning*. Springer Singapore, 2019, vol. 1037. [Online]. Available: [http://dx.doi.org/10.1007/978-981-13-9187-3\\_64](http://dx.doi.org/10.1007/978-981-13-9187-3_64)
- [20] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets," *Information Systems Frontiers*, 2021.
- [21] U. Kumari, A. K. Sharma, and D. Soni, "Sentiment analysis of smart phone product review using SVM classification technique," in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. IEEE, aug 2017, pp. 1469–1474. [Online]. Available: <https://ieeexplore.ieee.org/document/8389689/>
- [22] G. Cho, J. Yim, Y. Choi, J. Ko, and S. H. Lee, "Review of machine learning algorithms for diagnosing mental illness," *Psychiatry Investigation*, vol. 16, no. 4, pp. 262–269, 2019.
- [23] I. Syarif, N. Ningtias, and T. Badriyah, "Study on Mental Disorder Detection via Social Media Mining," *2019 4th International Conference on Computing, Communications and Security, ICCCS 2019*, pp. 1–6, 2019.
- [24] D. V. Devi, C. K. Kumar, and S. Prasad, "A Feature Based Approach for Sentiment Analysis by Using Support Vector Machine," *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, pp. 3–8, 2016.
- [25] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," pp. 511–520, 2018.
- [26] M. Srividya, S. Mohanavalli, and N. Bhalaji, "Behavioral Modeling for Mental Health using Machine Learning Algorithms," *Journal of Medical Systems*, vol. 42, no. 5, 2018.
- [27] MINSA Ministerio de Salud Perú, "Plan de Salud Mental 2020-2021," *Minsa*, vol. 2021, p. 60, 2020. [Online]. Available: <http://bvs.minsa.gob.pe/local/MINSA/5092.pdf>
- [28] IPSOS, "GLOBAL HEALTH MONITOR 2021," no. October, p. 34, 2021. [Online]. Available: [www.ipsos.com](http://www.ipsos.com)

# Modeling and Predicting Blood Flow Characteristics through Double Stenosed Artery from Computational Fluid Dynamics Simulations using Deep Learning Models

Ishat Raihan Jamil<sup>1</sup>

Department of Mechanical Engineering  
Bangladesh University of Engineering and Technology  
Dhaka, Bangladesh

Mayeesha Humaira<sup>2</sup>

Department of Computer Science and Engineering  
Ahsanullah University of Science and Technology  
Dhaka, Bangladesh

**Abstract**—Establishing patient-specific finite element analysis (FEA) models for computational fluid dynamics (CFD) of double stenosed artery models involves time and effort, restricting physicians' ability to respond quickly in time-critical medical applications. Such issues might be addressed by training deep learning (DL) models to learn and predict blood flow characteristics using a dataset generated by CFD simulations of simplified double stenosed artery models with different configurations. When blood flow patterns are compared through an actual double stenosed artery model, derived from IVUS imaging, it is revealed that the sinusoidal approximation of stenosed neck geometry, which has been widely used in previous research works, fails to effectively represent the effects of a real constriction. As a result, a novel geometric representation of the constricted neck is proposed which, in terms of a generalized simplified model, outperforms the former assumption. The sequential change in artery lumen diameter and flow parameters along the length of the vessel presented opportunities for the use of LSTM and GRU DL models. However, with the small dataset of short lengths of doubly constricted blood arteries, the basic neural network model outperforms the specialized RNNs for most flow properties. LSTM, on the other hand, performs better for predicting flow properties with large fluctuations, such as varying blood pressure over the length of the vessels. Despite having good overall accuracies in training and testing across all the properties for the vessels in the dataset, the GRU model underperforms for an individual vessel flow prediction in all cases. The results also point to the need of individually optimized hyperparameters for each property in any model rather than aiming to achieve overall good performance across all outputs with a single set of hyperparameters.

**Keywords**—Double stenosed artery; CFD; neural network; LSTM; GRU

## I. INTRODUCTION

Cardiovascular diseases (CVDs) are the most common causes of death around the world. Heart attacks are typically sudden occurrences caused by the narrowing and blockage of blood vessels<sup>1</sup>. A stenosed artery refers to the narrowing of a blood vessel caused by the deposition of atherosclerotic plaque on the inner walls of the arterial lumen [2, 3]. These cholesterol

and fatty deposits lead to a swollen and inflamed inner arterial wall which restricts the flow of oxygenated blood cells, nutrients, and other essential substances from reaching the heart muscles [4, 5]. Cholesterol have been shown to accelerate the formation of plaque in arteries, eventually obstructing the bloodstream and altering hemodynamics [6]. When the plaque ruptures, the accumulated fatty acids, platelets, and dead cells may coagulate, resulting in thrombosis formation [7]. In the case of a coronary or cerebral artery, the blood clot may have fatal consequences since it will also cut off the blood flow to the cerebral region of the brain or the myocardial heart wall [8]. The likelihood of developing thrombosis is highly dependent on the thickness of the plaque, the characteristics of infected blood, and blood pressure [9].

Numerous studies observed pulsing flow behavior and constant dampening of its oscillations, which they attributed to the flexibility of blood vessels [10]. Coronary artery disorder is critical in hemodynamics because it alters the flow pattern, resulting in variations in the wall pressure and shear stress of the arteries. As a result, health researchers must determine the flow velocity and amount of shear stress in arteries. A substantial part of the published research [11] focused on the physiological origins of the disease as they relate to blood vessels. However, few have made strides in understanding the underlying physics of the illness in order to better understand the cause and, as a result, paving the way to less invasive and more long-term treatments. Medical imaging can be utilized to visualize the areas of fatty deposits inside artery walls, but it is not capable of providing numerical data in the same way that computational fluid dynamics (CFD) simulations are capable of providing. According to Kompatsiaris et al. [12] and Liu et al. [13], computational simulations can offer an in-depth evaluation of flow resistance owing to wall shear stress ( $WSS$ ) on blood vessel walls, blood flow rates, and pressure changes. CFD results generated by modeling vessels in the relevant regions may be compared to the reliability of mathematical data. It is possible to develop a less invasive dependable method for medical diagnosis by integrating physician expertise with data derived from realistic computational fluid dynamics models. Owing to the vessel's small dimensions, in vitro, and in vivo flow field experiments are not representative and accurate. Thus, with improved software development and computer

<sup>1</sup><https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>

efficiency, CFD may replace such experimental approaches. CFD has been already used in several studies involving blood flow through vessels. Fazlay et al. [14] used CFD to show that following a double stenosed region in an artery, blood flow is hampered significantly at maximum systolic velocity and acceleration. Jianhuang et al. [15] coupled transient blood flow with elastic artery to evaluate unsteady flow characteristics along its length using computational fluid dynamics. Mukesh et al. [1] used an in-house CFD solver to verify and simulate blood flow via a stenosed artery. Mehdi et al. [17] compared several turbulent models from blood flow through vessels and concluded that inaccuracies are introduced by assuming the flow to be laminar.

Setting up patient-specific finite element analysis (FEA) models for CFD takes time and effort, limiting quick response to physicians in time-sensitive medical applications. As such, Liang et al. [18] created deep learning (DL) algorithm to predict aortic stress distributions. The DL model was developed to use FEA data and directly produce aortic wall stress distributions, skipping the FEA calculation step entirely. Arzani et al. [19] proposed a Physics-informed neural networks framework for predicting near-wall blood flow and wall shear stress from sparse velocity data concentrated in an interest region. Gao et al. [20] proposed a deep neural network approach that allows machines to recognize fractional flow reserve values directly from static coronary CT angiography images. Such progressions in computer science open up new horizons for further development. For instance, if the diameters of the stenosed aortic vessel are considered at regular intervals, a special kind of artificial neural network called recurrent neural network (RNN) can be implemented. RNN's internal memory allows them to comprehend sequential data. Their ability to retain crucial details about the preceding step, such as aortic diameter, could enable them to predict occurrences in the next step such as *WSS*, blood pressure or flow velocity more accurately. Previously RNN has been successfully implemented for malware classification [21], 3D shape generation [22], traffic forecasting [23], and speech enhancement [24]. While former research works explored various aspects of blood flow through arteries and investigated the predictive capabilities of several DL models, they overlooked such sequential trends in the variation of aortic diameter that occurs within the vessels. As such, it is unclear what effects these specialized RNNs might have on the fast prediction of blood flow characteristics through the arteries. A double stenosed artery is of particular interest in this study since it not only provides wide variations in aortic diameters within a short length but also poses serious health hazards within the human body, so much so that stenting might be required. Thus predicting the flow within them quickly might aid medical researchers with stent improvements and deployment. However, there are several hurdles that must be addressed before such analysis can be performed. As a result, the paper is divided into the following sections in order to discuss them further:

- Section II explains the computational fluid dynamics simulations.
- Section III describes the deep learning models utilized in this study.
- Section IV demonstrates the data source and the organization of the dataset.

- Section V presents the hyperparameters that were fine-tuned in this study.
- Section VI discusses the results obtained.
- Section VII presents the conclusion that can be drawn from the findings.

## II. COMPUTATIONAL FLUID DYNAMICS SIMULATIONS

Several data are necessary to construct a dataset in order to implement artificial intelligence (Ai). The lack of sufficient medical data on blood flow patterns in doubly stenosed arteries necessitates the use of computational approaches to generate the data. This allows for the exploration and visualization of the variations in blood flow behavior induced by several combinations of stenosis at varying distances apart. This section describes the processes used to set up CFD simulations and compares a couple of simplified models to identify which one best depicts the real flow characteristics, using an actual double stenosed artery model as a reference.

### A. Governing Equations

If the Navier–Stokes equation is interpreted as the sum of an average and an oscillating component for each variable, then the continuity and Reynolds averaged Navier–Stokes equations (RANS) are as follows:

$$\frac{\partial \bar{u}_i}{\partial x_i} = 0 \quad (1)$$

$$\frac{D\bar{u}_i}{Dt} = -\frac{1}{\rho} \frac{\partial \bar{P}}{\partial x_i} + \frac{\partial}{\partial x_j} \left( \frac{\mu + \mu_T}{\rho} \left( \frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i} \right) \right) \quad (2)$$

where  $\mu$  represents the summation of laminar  $\mu_0$  and turbulent  $\mu_T$  viscosities:

$$\mu = \mu_0 + \mu_T \quad (3)$$

For  $K$ - $\varepsilon$  standard turbulence model,  $\mu_T$  is computed as:

$$\mu_T = \rho c_\rho \frac{K^2}{\varepsilon} \quad (4)$$

where  $K$  represents the turbulence kinetic energy and  $\varepsilon$  denotes the rate of turbulence dissipation.

A straight double stenosed artery can be simply modeled as a tube with a diameter  $D$  with stenosed necks  $S_1$  and  $S_2$  separated by a distance of  $L$ . The degree of obstruction of the stenosed regions can be expressed as follows:

$$\%S = \frac{D-d}{D} \times 100\% \quad (5)$$

where  $d$  is the lumen diameters at neck  $S$ . The fraction of lumen opening at the neck then can be addressed as:

$$\text{Fraction of lumen opening} = 1 - \%S/100 \quad (6)$$

### B. Simulation Setup

Solidworks is used to model the arteries for this study. The blood vessels are assumed to be rigid with the no-slip condition at the arterial wall. Ansys Fluent software is used to set up the simulation and solve the RANS equations utilizing the finite volume method (FVM). The second-order upwind scheme was employed to spatially discretize the governing equations and the SIMPLE method was used to manage the pressure-velocity decoupling [16]. Blood is considered to be an incompressible fluid with a density of  $1050 \text{ kg/m}^3$  and a viscosity of  $0.0033 \text{ Pa}\cdot\text{s}$  [17]. The inlet flow velocities are obtained from the velocity profile presented by Fazlay et al. [14]. The authors pointed out five particular velocities from the waveform:  $0.21 \text{ m/s}$ ,  $0.33 \text{ m/s}$ ,  $0.28 \text{ m/s}$ ,  $0.14 \text{ m/s}$ , and  $0.09 \text{ m/s}$  at maximum systolic acceleration, systolic velocity, systolic deceleration, diastolic velocity, and at minimum systolic velocity respectively. Due to the presence of plasma, platelets, and suspended cells, blood has the characteristics of non-Newtonian fluid [16]. However, numerous previous CFD research treated blood as a Newtonian fluid [14, 17, 25]. In fact, Ku et al. [26] observed that for Reynolds numbers ( $Re$ ) ranging from 110 to 850 in big arteries, the non-Newtonian impact of blood is insignificant. As such, blood is deemed Newtonian in this analysis since  $Re$  stays within this range. The simulations are carried out with a time step of  $0.0001 \text{ s}$  and a mesh element size of  $0.112 \text{ mm}$ .

Fig. 1 shows the velocity profile for the sinusoidal stenosed artery model presented by Fazlay et al. [14] at a distance equal to the vessel's diameter  $D$  away from the stenosed neck using the aforementioned blood flow characteristics and the  $K-\epsilon$  standard turbulence model. The velocity curve closely matches the velocity profile of the model suggested by Mehdi et al. [17] and has a good agreement with the experimental results of Ahmed and Giddens [27]. As a result, further simulations are performed using this particular turbulence model.

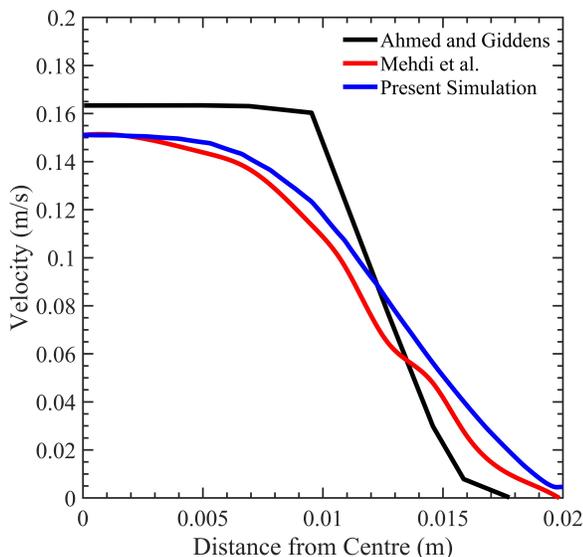


Fig. 1. Validation Test Comparing Velocity Profile of Present Simulation with Simulation Result of Mehdi et al. and Experimental Results of Ahmed and Giddens at a Distance Equal to the Vessel's Diameter  $D$  Away from the Stenosed Neck.

### C. Modeling Double Stenosed Artery

Fig. 2 illustrates the modeling of a straight section of an actual doubled stenosed artery obtained via intravascular ultrasound (IVUS) imaging using a  $3 \text{ fr}$  catheter and a  $1 \text{ mm}$  guidewire, with a pullback rate of  $1 \text{ mm/s}$ . It has an average non-stenosed hydraulic diameter of approximately  $4 \text{ mm}$ , with  $40.25\%$  and  $32\%$  stenosis situated  $10 \text{ mm}$  apart. Fig. 3 shows two simplified representations of the actual artery with similar stenoses. Unlike the sinusoidal equation-generated model [14] for a similar configuration, the actual model exhibits a gradual decrease in lumen diameter, as visible from Fig. 4, while Fig. 4 illustrates the variation in blood flow patterns through them. This simplification has a significant effect on the flow characteristics of blood. As demonstrated by Fig. 4, the steep sinusoidal stenosed edges exhibit a wide variation in average velocity ( $V_{avg}$ ), wall shear stress ( $WSS$ ), and pressure from those obtained from the actual model simulation for input velocity of  $0.3 \text{ m/s}$ . As such, a more representative model is required. Fig. 3 presents another model denoted as the splined model. This model features a  $25\%$  stenosis region  $5 \text{ mm}$  upstream and  $5 \text{ mm}$  downstream of the main stenosed neck. When such circular cross-sections are joined with spline guidelines, the simplified model captures the actual model's naturally formed gradual stenosis characteristics. This results in a  $WSS$  curve that is more similar to that of the actual artery model as can be seen in Fig. 4. The  $V_{avg}$  curve also shows a similar trend but is visibly higher due to the absence of surface irregularities to retard the flow as in the actual model. On the other hand, both simplified models fail to represent the actual model's pressure fluctuations effectively, especially at the diverging sections following the stenosed necks.

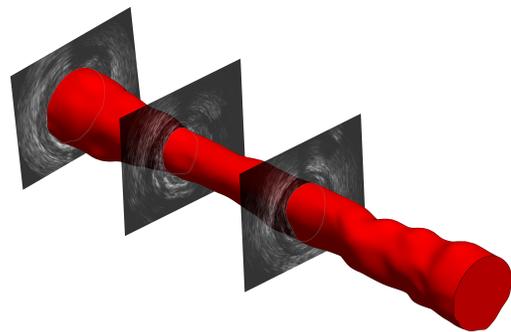


Fig. 2. Modeling of an Actual Double Stenosed Artery from IVUS Images. For Clarity, Only a Few IVUS Images are Shown.

Fig. 5 illustrates the aforementioned flow characteristics visually. The graphic clearly illustrates the influence of naturally produced uneven surfaces on the actual model. Initially, the blood pressure is rather high in all three models. As the blood reaches the first stenosed neck, the pressure gradually drops to or below zero. According to Bernoulli's principle, the decrease in pressure induces an increase in flow velocity in the stenosed region, as also visible from the figure. This results in a significant rise in  $WSS$  at the constriction site. According to the

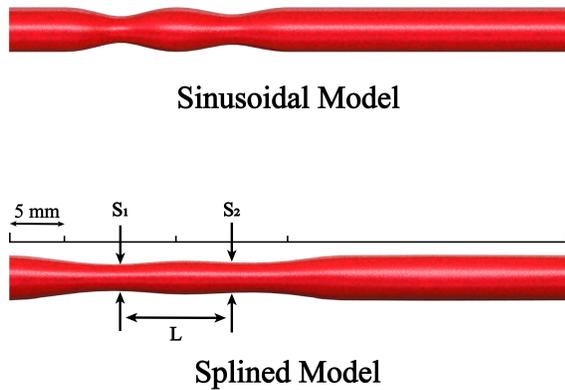


Fig. 3. Comparison Between the Sinusoidal and Splined Representation of Stenosed Neck Geometry in a Simplified Model of the Actual Double Stenosed Artery.

continuity equation, when blood travels further into the first diverging section, the increase in the cross-sectional area results in a reduction in flow velocity. As a consequence, pressure rises and  $WSS$  falls. As the blood approaches the next stenosed neck, the velocity increases again but to a lower magnitude due to less constriction there. The pressure decreases once again, along with a slight rise in  $WSS$ . Further downstream, the velocity attains equilibrium, and both pressure and  $WSS$  approach near-zero values. Although not perfect, Fig. 4 and Fig. 5 indicate that the splined model would be a better match than the sinusoidal model for the artificial intelligence (Ai) implementation in the next section. Such is the cost of creating a generalized simplified model without being patient-specific. From these discussions, it is also apparent that the progressive change in the cross-sectional diameter of the stenosed aortic vessel gradually affects blood flow characteristics. This enables the generation of a sequential dataset by taking into account the diameters and flow properties at regular intervals along the length of the blood vessel. The mesh independence test, as illustrated in Fig. 6, demonstrates that changing the mesh element size has a minor influence on the flow characteristics of blood through the splined double stenosed artery model. Between  $0.099\text{ mm}$  and  $0.112\text{ mm}$ , the variation in simulation results is significantly less. As such, it is more reasonable to adopt a mesh element size of roughly  $0.112\text{ mm}$  throughout this study to achieve high CFD simulation accuracy without being too computationally expensive.

### III. DEEP LEARNING MODELS

A recurrent neural network (RNN) is a special kind of Ai network with internal memory that enables it to comprehend sequential data. However, the basic RNN is afflicted by a phenomenon known as vanishing gradient [28]. Long short-term memory (LSTM) and gated recurrent units (GRU) are special kinds of RNN networks developed to mitigate the problem. Their capacity to retain crucial details from the preceding step, such as aortic diameter, would allow them to effectively forecast occurrences in the following step, such as wall shear stress ( $WSS$ ), average velocity ( $V_{avg}$ ) of blood, and pressure. Three techniques are employed in this study to

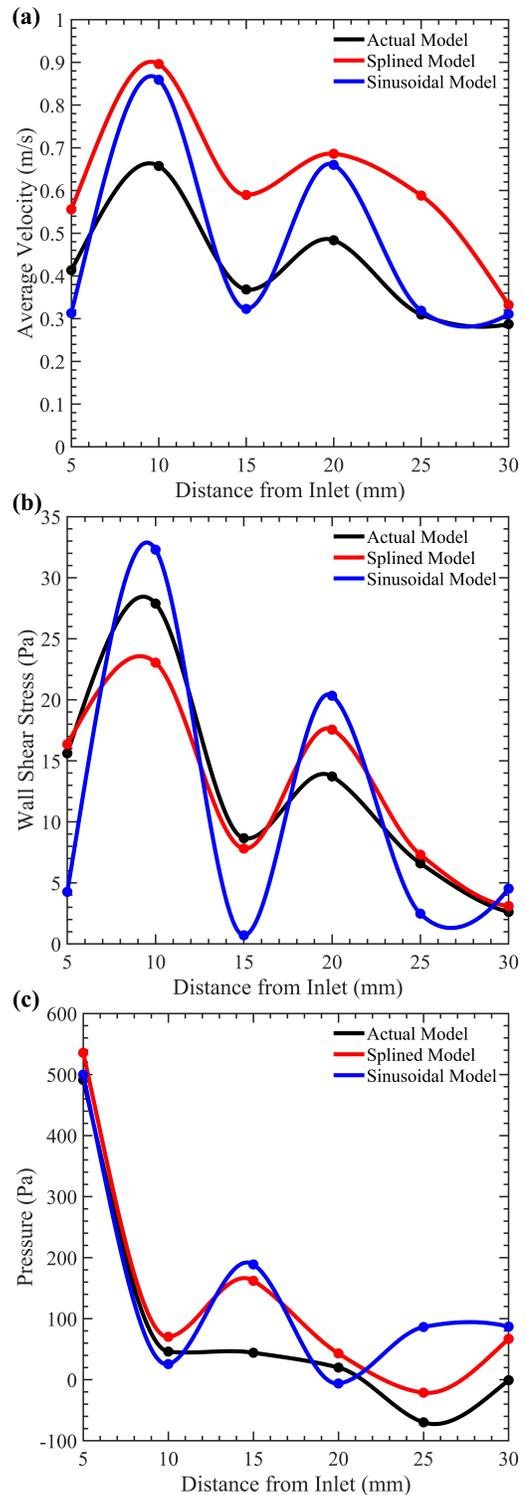


Fig. 4. Comparison of Flow Characteristics Through the Actual and Simplified Double Stenosed Artery Models.

predict these flow properties: Gated Recurrent Unit (GRU), Long short-term memory (LSTM), and Neural Network (NN) models. All three models used inlet velocity and percentage lumen openings at eleven locations along the  $50\text{ mm}$  long blood artery at regular  $5\text{ mm}$  intervals to predict the blood

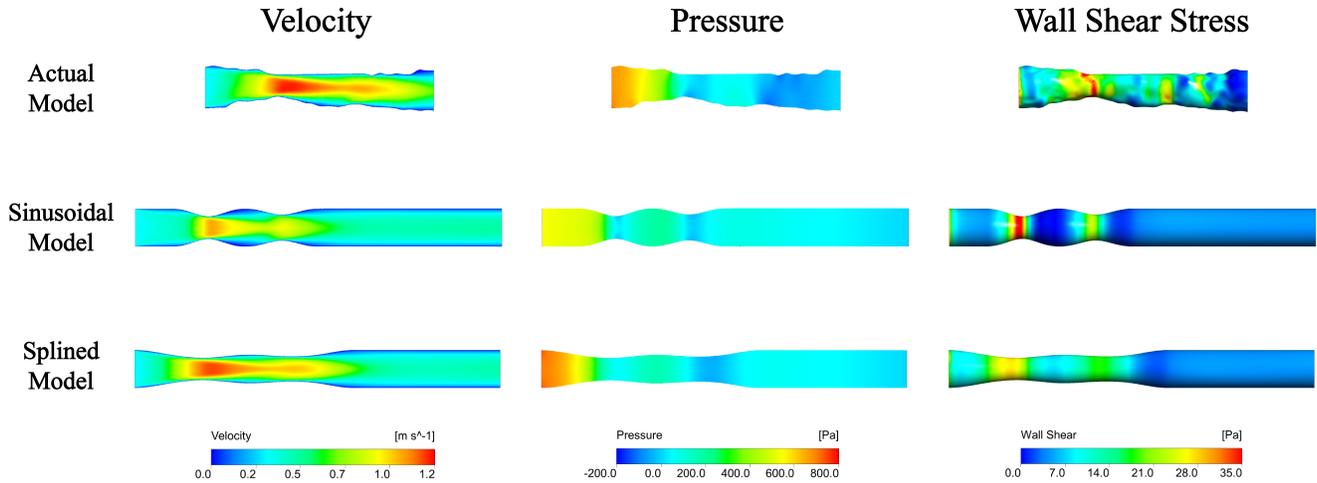


Fig. 5. Visualization of Flow Properties Through the Actual and Simplified Double Stenosed Artery Models.

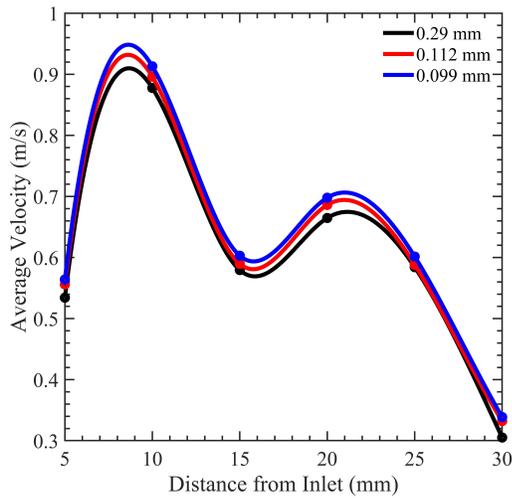


Fig. 6. Mesh Independence Test with Various Sizes of Mesh Elements.

flow characteristics at those positions. This section highlights each of the three deep learning models.

#### A. Gated Recurrent Unit Model

Gated Recurrent Unit (GRU) [29] is a special kind of recurrent neural network that consists of an update gate and a reset gate. GRU's update gate determines how much data from previous units must be passed on. The update gate computes  $z_t$  for time step  $t$  using the formula:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (7)$$

where  $z_t$  is update gate output at the current timestamp,  $W_z$  is weight matrix at update gate,  $h_{t-1}$  information from previous units, and  $x_t$  is input at the current unit.

The model used the reset gate to determine how much information from previous units should be erased. This is

calculated using the following formula:

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (8)$$

where  $r_t$  is reset gate output at current timestamp,  $W_r$  is weight matrix at reset gate,  $h_{t-1}$  information from previous units, and  $x_t$  is input at the current unit. The relevant data from earlier units were stored in the current memory content using this formula:

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (9)$$

where  $h_t$  is current memory content,  $W$  is weight at current unit,  $r_t$  is reset gate output at current timestamp,  $h_{t-1}$  is information from previous units, and  $x_t$  is input at the current unit.

Final memory at the current unit was a vector that stored and conveyed the current unit's final information to the next layer. This was computed using the following formula:

$$h_t = (1 - z_t) * h_{t-1} + z_t \tilde{h}_t \quad (10)$$

where  $h_t$  is final memory at the current unit,  $z_t$  is update gate output at current timestamp,  $h_{t-1}$  is information from previous units, and  $h_t$  is current memory content.

#### B. Long Short-term Memory Model

Another sort of RNN is the Long short-term memory (LSTM) [30]. In contrast to the GRU, the LSTM contains three gates: the forget gate, the update gate, and the output gate. The LSTM gates' formulae are as follows:

$$i_t = \sigma(W_i [h_{t-1}, x_t] + b_i) \quad (11)$$

$$f_t = \sigma(W_f [h_{t-1}, x_t] + b_f) \quad (12)$$

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \quad (13)$$

where  $i_t$  represents input gate,  $f_t$  represents forget gate,  $o_t$  represents output gate,  $\sigma$  represents sigmoid function,  $W_x$  represents weight of the respective gate( $x$ ) neurons,  $h_{t-1}$  represents output of previous LSTM block at timestamp  $t - 1$ ,  $x_t$  represents input at current timestamp and  $b_x$  represents biases for the respective gates( $x$ ).

Both GRU and LSTM models utilized the many-to-many combination, with an 11 node dense layer as the output, to predict the  $V_{avg}$ ,  $WSS$ , or blood pressure at the eleven positions, taking the aforementioned inputs. These models are depicted in Fig. 7. The hyperparameters of these models were varied in order to maximize the prediction accuracies across all three flow properties.

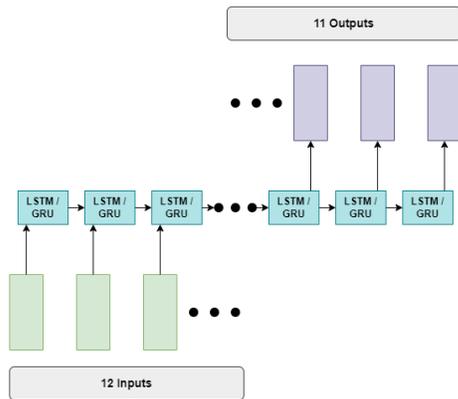


Fig. 7. Illustration of LSTM / GRU Ai Models in Many-to-Many Configuration.

### C. Neural Network Model

The other aspect of the present study is the neural network (NN) as shown in Fig. 8. It is defined as a collection of algorithms that are capable of correctly recognizing the underlying connections between a set of data via a method that replicates the way the human brain works. They are not limited to sequential data and are composed of nodes with assigned weights. Through the forward and backward propagation processes using labeled data, the network is able to fine-tune the weights to make accurate predictions. It also, applied the same inputs as the previous two models to predict the same flow features. The hyperparameters were also varied for this model to improve its accuracy.

## IV. DATASET

Since the present study involves a newly proposed doubled stenosed artery model, a dataset of CFD simulation results relating to it is unavailable. As such, a custom dataset containing 180 data points was constructed. To create the dataset, several configurations of fractions of lumen opening at each stenosed neck, gaps between them, and inlet velocities, as previously mentioned in the simulation setup section, were used. In particular, stenosis severity levels of 25%, 50%, and 75% were applied at individual necks, with 10 mm, 15 mm, 20 mm, and 25 mm spacing between them. 90% of the total data were utilized for training, while the remaining 10% was

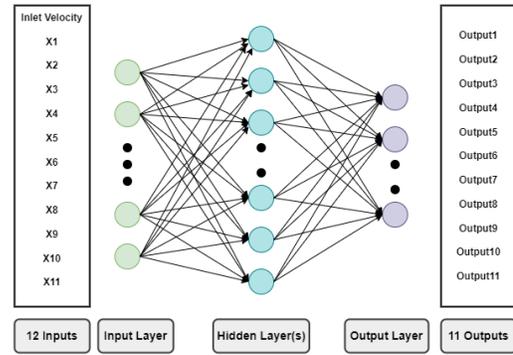


Fig. 8. Illustration of Neural Network (NN) Ai Model.

used for validation. A different test set containing 18 datapoints was also constructed using configurations that were absent in the main dataset, such as different inlet velocities and stenosis severity levels of 70%, 60%, 40%, 30%, etc. Due to the lack of additional IVUS images with similar double stenosis conditions, the blood vessel employed is a generalized form of a patient-specific actual model. Overcoming such hurdles, as well as including CFD simulations of curved vessels in the future could make the dataset even more beneficial.

## V. HYPERPARAMETERS

In order to improve the accuracy of these three models, several hyperparameters were tuned. Table I summarizes the ranges of the parameters varied. Firstly, different units for the LSTM / GRU models and different numbers of hidden layers (HL), containing 12 nodes in each, for the NN model were tested to determine the conditions that performed well for predicting all three flow properties. Subsequently, the number of epochs and the learning rates were optimized to maximize the accuracies of the deep learning models. Other parameters such as the loss function, activation functions, and optimizer were kept uniform in all models to provide a fair comparison. Initially, min-max normalization was used to ensure that all of the data in the dataset was within the range of 0 to 1.

To compute the loss, each model utilized the Mean squared error (MSE) function which is represented by the formula:

$$L(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^N (y - \hat{y}_i)^2 \quad (14)$$

where  $\hat{y}$  is the predicted value,  $N$  is the number of data points, and  $y$  is the observed value. MSE in particular can penalize large errors more than smaller ones, making it a good choice for achieving multiple accurate predictions. Each dense layer utilized the sigmoid activation function, which produces a probabilistic output that exists exclusively between 0 and 1, following the equation:

$$\phi(z) = \frac{1}{1 + e^{-z}} \quad (15)$$

For its capability of achieving excellent results quickly and effectively, the Adam [31] optimizer was implemented in all of the models.

TABLE I. HYPERPARAMETERS ADAPTED IN DIFFERENT MODELS

|                          | GRU        | LSTM       | Neural Network |
|--------------------------|------------|------------|----------------|
| Units / Hidden Layer(HL) | units: 12  | units: 12  | HL: 1          |
|                          | units: 48  | units: 48  | HL: 2          |
|                          | units: 84  | units: 84  | HL: 3          |
|                          | units: 120 | units: 120 | HL: 4          |
| Epochs                   | 50000      | 50000      | 50000          |
|                          | 100000     | 100000     | 100000         |
|                          | 150000     | 150000     | 150000         |
|                          | 200000     | 200000     | 200000         |
| Learning Rate            | 0.01       | 0.01       | 0.01           |
|                          | 0.001      | 0.001      | 0.001          |
|                          | 0.0001     | 0.0001     | 0.0001         |
|                          | 0.00001    | 0.00001    | 0.00001        |

## VI. RESULT AND DISCUSSION

Training and testing accuracies were used to assess the three suggested Deep Learning approaches. These accuracies indicated the degree to which each model could correctly predict  $V_{avg}$ ,  $WSS$ , or blood pressure during testing or training. The variance in the accuracies for varying the number of units in the GRU and LSTM models, as well as for different the number of hidden layers in the NN model is represented in Table II. In these cases, the number of epochs and learning rate were kept constant at 100000 and 0.0001 respectively. Owing to a small dataset, the LSTM model with 84 units overall performed well across all flow properties for both training and test sets, as visible in the table. In contrast, the GRU model did well throughout both sets with only 48 units. Extending the units beyond these ideal values seems to degrade the efficacy of both models. On the other hand, the NN model shows an overall decreased effectiveness in predicting the flow properties for both sets as compared to the other two models with the same number of epochs and learning rate. Nonetheless, testing with different numbers of hidden layers reveals that the smaller number of data prefers a shallower architecture with 2 intermediate layers only, having good accuracies on most occasions. Pressure values in the dataset were either very large or very small at certain points along the length of the vessel, as also visible from Fig. 4. As a consequence, when these numbers were normalized, tiny values were transformed to near-zero values, while bigger values were changed to near-one values. As such, each model only needed to learn either of these two extreme quantities at certain locations, leading to much higher accuracies for both training and test sets.

To further tune the hyperparameters and improve the accuracies of the models, the number of epochs was varied from

50000 to 200000 while keeping the learning rate constant at 0.0001. Based on the previous evaluations, the number of units for LSTM, GRU, and the number of hidden layers for the NN model is set to 84, 48, and 2 respectively. The results of these evaluations are reported in Table III. It can be seen that for the LSTM model, the highest accuracy for  $V_{avg}$  is obtained at 200000 epochs, whereas pressure achieves the greatest accuracy at 150000 epochs. The predicting effectiveness for  $WSS$  remains unchanged from 100000 to 200000 epochs. However, the average testing accuracies across three properties are much lower above and below 100000 epochs, indicating the model cannot generalize well in those ranges. The GRU on average also performs very well at 100000 epochs using both sets of data. The training dataset for the NN model clearly prefers the higher epochs but the peak average test set accuracy at 150000 epochs indicates non generalizing effect beyond this value.

Finally, the learning rates of each model are tuned, setting the other hyperparameters to their predetermined optimum values. The results of these investigations are shown in Table IV. The table suggests that LSTM performs very well in predicting the flow properties for learning rate in between 0.01 to 0.0001, whereas GRU is most accurate for a rate of 0.001. NN model on the other hand prefers a learning rate between 0.001 and 0.0001. In certain cases the larger learning rate overshoots, destabilizes the training process, and fails to reach optimum accuracy, for example, NN for predicting pressure of the flow. It also tends to overfit the data from both LSTM and NN models. Too small learning rate isn't beneficial either since it also lowers the overall accuracies of all the models across the three flow properties. Although not the most accurate in every instance, a learning rate of 0.001 that can generalize well is preferred by all three models to reasonably predict each of the flow properties from both datasets.

Fig. 9 compares the predicted flow characteristics obtained with the optimized hyperparameters to the CFD simulation results obtained with the real and splined models for the identical configuration with 40.25% and 32% stenosis located 10 mm apart. In the case of average velocity, the LSTM model overestimates, while the GRU model underestimates, and therefore fails to effectively estimate the flow pattern. On the other hand, the NN model is far more accurate at forecasting flow velocity patterns and is comparable to the simulation results achieved with the splined model. Again, for  $WSS$ , the NN models perform well and closely track the simulation outcomes. Interestingly, the prediction imperfections put the LSTM model's  $WSS$  pattern prediction closer to the actual model's simulation outcomes without even training on that model. Nevertheless, both LSTM and GRU models fail to predict the  $WSS$  pattern entirely. Then again, the LSTM model performs much better at predicting pressure fluctuations, while the GRU model's simplicity prevents it from learning the pressure changes effectively from the small dataset. Unlike the other two properties, the NN model seems to be less capable of comprehending flow pressure variation. Thus, Fig. 9 further demonstrates that tuning the hyperparameters to generalize the models for predicting all flow properties leads to their overall diminished performance. As such, although more effort is necessary, optimizing the models based on each of the flow characteristics individually would be more beneficial. Additionally, the graphic illustrates the NN model's supremacy

TABLE II. ACCURACIES ACHIEVED BY DIFFERENT MODELS BY VARYING THE NUMBER OF UNITS FOR THE GRU AND LSTM MODEL AND CHANGING THE NUMBER OF HIDDEN LAYERS FOR THE NEURAL NETWORK MODEL

| Model | Units/HL | Epoch  | Learning Rate | Training  |        |          | Testing   |        |          |
|-------|----------|--------|---------------|-----------|--------|----------|-----------|--------|----------|
|       |          |        |               | $V_{avg}$ | WSS    | Pressure | $V_{avg}$ | WSS    | Pressure |
| LSTM  | 12       | 100000 | 0.0001        | 0.8500    | 0.8778 | 0.8278   | 0.7778    | 0.7222 | 0.8889   |
|       | 48       |        |               | 0.8611    | 0.9833 | 0.9722   | 0.8333    | 0.7778 | 0.7778   |
|       | 84       |        |               | 0.8778    | 0.9778 | 0.9889   | 0.8889    | 0.8889 | 0.8889   |
|       | 120      |        |               | 0.9111    | 0.9722 | 0.9444   | 0.9444    | 0.7222 | 0.8333   |
| GRU   | 12       |        |               | 0.7944    | 0.7667 | 0.9500   | 0.8333    | 0.7222 | 0.9444   |
|       | 48       |        |               | 0.9000    | 0.9722 | 0.9722   | 0.9444    | 0.7778 | 0.9444   |
|       | 84       |        |               | 0.8833    | 0.9556 | 0.9556   | 0.7778    | 0.8333 | 0.9444   |
|       | 120      |        |               | 0.8389    | 0.9667 | 0.9611   | 0.9444    | 0.8889 | 0.9999   |
| NN    | 1        |        |               | 0.8444    | 0.9500 | 0.7167   | 0.8333    | 0.8333 | 0.7222   |
|       | 2        |        |               | 0.8222    | 0.9722 | 0.9389   | 0.8333    | 0.7778 | 0.9999   |
|       | 3        |        |               | 0.8222    | 0.9667 | 0.8778   | 0.7778    | 0.7222 | 0.9999   |
|       | 4        |        |               | 0.8167    | 0.8278 | 0.9333   | 0.8889    | 0.5556 | 0.9999   |

TABLE III. ACCURACIES ACHIEVED BY DIFFERENT MODELS BY VARYING THE NUMBER OF EPOCHS

| Model | Units/HL | Epoch  | Learning Rate | Training  |        |          | Testing   |        |          |
|-------|----------|--------|---------------|-----------|--------|----------|-----------|--------|----------|
|       |          |        |               | $V_{avg}$ | WSS    | Pressure | $V_{avg}$ | WSS    | Pressure |
| LSTM  | 84       | 50000  | 0.0001        | 0.8667    | 0.9389 | 0.7000   | 0.7778    | 0.7222 | 0.5000   |
|       |          | 100000 |               | 0.8778    | 0.9778 | 0.9889   | 0.8889    | 0.8889 | 0.8889   |
|       |          | 150000 |               | 0.8944    | 0.9778 | 0.9944   | 0.7778    | 0.7222 | 0.9999   |
|       |          | 200000 |               | 0.9222    | 0.9778 | 0.9722   | 0.7778    | 0.7778 | 0.7222   |
| GRU   | 48       | 50000  |               | 0.8389    | 0.9389 | 0.8222   | 0.8333    | 0.9444 | 0.7778   |
|       |          | 100000 |               | 0.9000    | 0.9722 | 0.9722   | 0.9444    | 0.7778 | 0.9444   |
|       |          | 150000 |               | 0.8667    | 0.9556 | 0.9667   | 0.8889    | 0.7778 | 0.9999   |
|       |          | 200000 |               | 0.8944    | 0.9667 | 0.9778   | 0.8889    | 0.8889 | 0.9999   |
| NN    | 2        | 50000  |               | 0.8500    | 0.9333 | 0.6444   | 0.8333    | 0.8333 | 0.6667   |
|       |          | 100000 |               | 0.8222    | 0.9722 | 0.9389   | 0.8333    | 0.7778 | 0.9999   |
|       |          | 150000 |               | 0.8222    | 0.9389 | 0.9778   | 0.8333    | 0.9444 | 0.9999   |
|       |          | 200000 |               | 0.8556    | 0.9778 | 0.9999   | 0.7778    | 0.8333 | 0.9999   |

TABLE IV. ACCURACIES ACHIEVED BY DIFFERENT MODELS BY VARYING THE LEARNING RATE

| Model | Units/HL | Epoch  | Learning Rate | Training  |        |          | Testing   |        |          |
|-------|----------|--------|---------------|-----------|--------|----------|-----------|--------|----------|
|       |          |        |               | $V_{avg}$ | WSS    | Pressure | $V_{avg}$ | WSS    | Pressure |
| LSTM  | 84       | 100000 | 0.01          | 0.8611    | 0.9944 | 0.9944   | 0.9444    | 0.7778 | 0.8889   |
|       |          |        | 0.001         | 0.9500    | 0.9778 | 0.9778   | 0.7778    | 0.8889 | 0.9999   |
|       |          |        | 0.0001        | 0.8778    | 0.9778 | 0.9889   | 0.8889    | 0.8889 | 0.8889   |
|       |          |        | 0.00001       | 0.8833    | 0.9500 | 0.500    | 0.8333    | 0.7222 | 0.6111   |
| GRU   | 48       | 100000 | 0.01          | 0.8389    | 0.9722 | 0.9778   | 0.8333    | 0.5556 | 0.9999   |
|       |          |        | 0.001         | 0.9333    | 0.9833 | 0.9833   | 0.9444    | 0.8333 | 0.9444   |
|       |          |        | 0.0001        | 0.9000    | 0.9722 | 0.9722   | 0.9444    | 0.7778 | 0.9444   |
|       |          |        | 0.00001       | 0.8611    | 0.9111 | 0.7778   | 0.8889    | 0.6667 | 0.8333   |
| NN    | 2        | 150000 | 0.01          | 0.8944    | 0.9833 | 0.3611   | 0.7778    | 0.6111 | 0.2778   |
|       |          |        | 0.001         | 0.8111    | 0.9944 | 0.9999   | 0.8889    | 0.8333 | 0.9999   |
|       |          |        | 0.0001        | 0.8222    | 0.9389 | 0.9778   | 0.8333    | 0.9444 | 0.9999   |
|       |          |        | 0.00001       | 0.7111    | 0.6111 | 0.4222   | 0.6111    | 0.5556 | 0.4444   |

in predicting short lengths of sequential data from a small dataset.

## VII. CONCLUSION

The present study explored the faithfulness of simplified models' flow characteristics to that of the actual model derived from IVUS imaging. The model with a sinusoidal representation of stenosed geometry, which has been widely used in earlier research, entirely fails to portray the actual model's fluctuations in flow property patterns. Although not fully perfect, owing to non-circular cross-sections of the actual model, the newly proposed splined model stands out as a better representation for the construction of a database with various percentage stenoses with varying gaps between them,

for implementation of artificial intelligence. The sequential nature of the input and subsequently the flow properties opened up opportunities for specialized RNNs to be implemented. As it turns out, the short lengths of the vessel and small dataset prefer a simpler, less sophisticated conventional neural network model with shallow architecture for efficiently predicting most flow parameters, such as average velocity and wall shear stress. On the other hand, the considerable variation in pressure along the short length of the vessel favors the computationally expensive LSTM model with a large number of units. The simpler GRU model, although generalized well in terms of over accuracies across both training and test sets, fails to generate better predictions than the other two Ai models for any individual double stenosed artery. This highlights the fact that

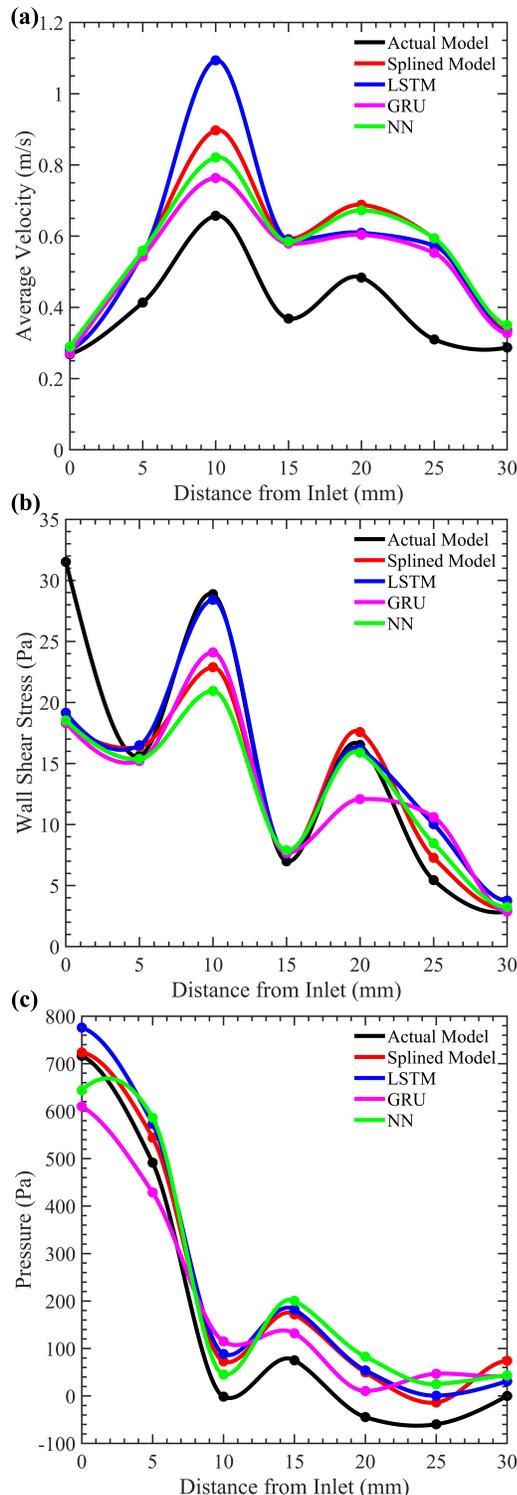


Fig. 9. Comparison Between Predicted Flow Characteristics with Optimized Hyperparameters and Simulated Results of Actual and Splined Artery Models.

instead of aiming to achieve overall good performance across all outputs with a single set of hyperparameters, each property needs to be addressed and models optimized individually. The major limitation of this study has been its small dataset. With more CFD simulations featuring stenosis configurations in

between those used in this training dataset, as well as with more inlet blood flow velocities within the relevant range it might be possible to improve the efficiencies of the Ai models. Future work on this subject might involve simulating and forecasting flow characteristics for curved vessels in order to make it an even better representation of the arteries naturally found in the human body. In addition, new Ai models might be developed to generate images directly from input fraction lumen opening data in order to better analyze blood flow patterns through double stenosed arteries. Such advancements, as well as the current findings, would allow medical researchers to swiftly estimate the severity of blood flow obstructions through constricted arteries, thereby assisting in stent development and deployment.

#### DECLARATION OF COMPETING INTEREST

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

#### ACKNOWLEDGMENT

The authors would like to express their gratitude to Dr. Md. Habibur Rahman, Associate Professor and Senior Consultant Cardiologist at National Heart Foundation Hospital & Research Institute, for providing medical expertise and IVUS imaging for this study.

#### REFERENCES

- [1] M. Roy, B. Singh Sikarwar, M. Bhandwal, and P. Ranjan, "Modelling of Blood Flow in Stenosed Arteries," *Procedia Comput. Sci.*, vol. 115, pp. 821–830, 2017, doi: 10.1016/j.procs.2017.09.164.
- [2] C. A. Taylor, T. A. Fonte, and J. K. Min, "Computational fluid dynamics applied to cardiac computed tomography for noninvasive quantification of fractional flow reserve: scientific basis," *Journal of the American College of Cardiology*, vol. 61(22), pp. 2233–41, 2013, doi: 10.1016/j.jacc.2012.11.083.
- [3] E. Geiringer, "Intimal vascularization and atherosclerosis," *The Journal of Pathology and Bacteriology*, vol. 63(2), pp. 201–211, 1951.
- [4] M. Thiriet, "Diseases of the Cardiac Pump," Cham: Springer International Publishing, vol. 7(1), pp. 2–20, 2015.
- [5] A. R. Lucas, R. Korol, and C. J. Pepine, "Inflammation in atherosclerosis: some thoughts about acute coronary syndromes," *Circulation*, vol. 113(17), pp. 728–32, 2006.
- [6] P. M. Ridker, M. J. Stampfer and, N. Rifai, "Novel risk factors for systemic atherosclerosis," *JAMA*, vol. 285(19), pp. 2481–2485, 2001.
- [7] R. Virmani, F. D. Kolodgie, A. P. Burke, A. Farb & S. M. Schwartz, "Lessons from sudden coronary death: a comprehensive morphological classification scheme for atherosclerotic lesions," *Atherosclerosis, Thrombosis, and Vascular Biology*, vol. 20(5), pp. 1262–1275, 2000.
- [8] P. Libby and, D. I. Simon, "Inflammation and thrombosis: the clot thickens," *Circulation*, vol. 103(13), pp. 1718–1720, 2001.
- [9] B. Furie, and B. C. Furie, "Mechanisms of thrombus formation," *The New England Journal of Medicine*, vol. 359(9), pp. 938–49, 2008.
- [10] J.R. Cebal, F. Mut, M. Raschi, E. Scrivano, R. Ceratto, P. Lylyk, and C.M. Putman, "Aneurysm rupture following treatment with flow-diverting stents: computational hemodynamics analysis of treatment," *Am. J. Neuroradiol.*, vol. 32, pp. 27–33, 2011.
- [11] D.N. Ku, "Blood flow in arteries," *Annu. Rev. Fluid Mech* 1997, vol. 29, pp. 399–436, 1997.
- [12] I. Kompatsiaris, D. Tzovaras, V. Koutkias, and M. G. Strintzis, "Deformable boundary detection of stents in angiographic images," *IEEE Trans. Med. Imaging*, vol. 19, no. 6, pp. 652–662, Jun. 2000, doi: 10.1109/42.870673.
- [13] G. Liu et al., "Numerical Simulation of Flow in Curved Coronary Arteries with Progressive Amounts of Stenosis Using Fluid-Structure Interaction Modelling," *J. Med. Imaging Health Inform.*, vol. 4, no. 4, pp. 605–611, Aug. 2014, doi: 10.1166/jmihi.2014.1301.

- [14] Md. F. Rubby, Md. S. Rana, and A. B. M. T. Hasan, "Hemodynamics of Physiological Blood Flow through a Double Stenosed Artery," *Procedia Eng.*, vol. 105, pp. 893–901, 2015, doi: 10.1016/j.proeng.2015.05.092.
- [15] J. Wu, G. Liu, W. Huang, D. N. Ghista, and K. K. L. Wong, "Transient blood flow in elastic coronary arteries with varying degrees of stenosis and dilatations: CFD modelling and parametric study," *Comput. Methods Biomech. Biomed. Engin.*, vol. 18, no. 16, pp. 1835–1845, Dec. 2015, doi: 10.1080/10255842.2014.976812.
- [16] A. Alshare, B. Tashtoush, and H. H. El-Khalil, "Computational Modeling of Non-Newtonian Blood Flow Through Stenosed Arteries in the Presence of Magnetic Field," *J. Biomech. Eng.*, vol. 135, no. 11, pp. 114503, Nov. 2013, doi: 10.1115/1.4025107.
- [17] M. Jahangiri, M. Saghafian, and M. R. Sadeghi, "Numerical Study of Turbulent Pulsatile Blood Flow through Stenosed Artery Using Fluid-Solid Interaction," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–10, 2015, doi: 10.1155/2015/515613.
- [18] L. Liang, M. Liu, C. Martin, and W. Sun, "A deep learning approach to estimate stress distribution: a fast and accurate surrogate of finite-element analysis," *J. R. Soc. Interface*, vol. 15, no. 138, pp. 20170844, Jan. 2018, doi: 10.1098/rsif.2017.0844.
- [19] A. Arzani, J. X. Wang, and R. M. D'Souza, "Uncovering near-wall blood flow from sparse data with physics-informed neural networks," arXiv preprint arXiv:2104.08249, 2021.
- [20] Z. Gao, X. Wang, S. Sun, D. Wu, J. Bai, Y. Yin, X. Liu, H. Zhang, and V.H.C. de Albuquerque, "Learning physical properties in complex visual scenes: an intelligent machine for perceiving blood flow dynamics from static CT angiography imaging," *Neural Networks*, vol. 123, pp.82-93, 2020.
- [21] B. Athiwaratkun and J. W. Stokes, "Malware classification with LSTM and GRU language models and a character-level CNN," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, pp. 2482–2486, Mar. 2017, doi: 10.1109/ICASSP.2017.7952603.
- [22] R. Wu, Y. Zhuang, K. Xu, H. Zhang, and B. Chen, "PQ-NET: A Generative Part Seq2Seq Network for 3D Shapes," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 826–835, Jun. 2020, doi: 10.1109/CVPR42600.2020.00091.
- [23] P. Sun, A. Boukerche, and Y. Tao, "SSGRU: A novel hybrid stacked GRU-based traffic volume prediction approach in a road network," *Comput. Commun.*, vol. 160, pp. 502–511, Jul. 2020, doi: 10.1016/j.comcom.2020.06.028.
- [24] L. Sun, J. Du, L.-R. Dai, and C.-H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in 2017 Hands-free Speech Communications and Microphone Arrays (HSCMA), San Francisco, CA, USA, 2017, pp. 136–140. doi: 10.1109/HSCMA.2017.7895577.
- [25] S. S. Shishir, Md. A. K. Miah, A. K. M. S. Islam, and A. B. M. T. Hasan, "Blood Flow Dynamics in Cerebral Aneurysm - A CFD Simulation," *Procedia Eng.*, vol. 105, pp. 919–927, 2015, doi: 10.1016/j.proeng.2015.05.116.
- [26] J. P. Ku, C. J. Elkins, and C. A. Taylor, "Comparison of CFD and MRI Flow and Velocities in an In Vitro Large Artery Bypass Graft Model," *Ann. Biomed. Eng.*, vol. 33, no. 3, pp. 257–269, Jan. 2005, doi: 10.1007/s10439-005-1729-7.
- [27] S. A. Ahmed and D. P. Giddens, "Pulsatile poststenotic flow studies with laser Doppler anemometry," *J. Biomech.*, vol. 17, no. 9, pp. 695–705, Jan. 1984, doi: 10.1016/0021-9290(84)90123-4.
- [28] M. Tanti, A. Gatt, and K. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?," pp. 51–60, 2018, doi: 10.18653/v1/w17-3506.
- [29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," arXiv preprint arXiv:1412.3555, 2014.
- [30] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [31] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc., pp. 1–15, 2015.

# NLI-GSC: A Natural Language Interface for Generating SourceCode

Aaqib Ahmed R.H. Ansari<sup>1</sup>  
Vidyalankar Institute of Technology  
University of Mumbai  
Mumbai, India

Dr. Deepali R. Vora<sup>2</sup>  
Symbiosis Institute of Technology  
Symbiosis International University  
Pune, India

**Abstract**—There are many different programming languages and each programming language has its own structure or way of writing the code, it becomes difficult to learn and frequently switch between different programming languages. Due to this reason, a person working with multiple programming languages needs to look at documentations frequently which costs time and effort. In the past few years, there have been significant increase in the amount of papers published on this topic, each providing a unique solution to this problem. Many of these papers are based on applying NLP concepts in unique configuration to get the desired results. Some have used AI along with NLP to train the system to generate source-code in specific language, and some have trained the AI directly without pre-processing the dataset with NLP. All of these papers face two problems: a lack of proper dataset for this particular application and each paper can convert natural language into only one specified programming language source-code. This proposed system shows that a language independent solution is a feasible alternate for writing source-code without having full knowledge about a programming language. The proposed system uses Natural Language Processing to convert Natural Language into programming language-independent pseudo code using custom Named Entity Recognition and save it in XML (eXtensible Markup Language) format which is an intermediate step. Then, using traditional programming, this system converts the generated pseudo code into programming language-dependent source-code. In this paper, another novel method has been proposed to create dataset from scratch using predefined structure that is filled with predefined keywords creating unique combination of training dataset.

**Keywords**—Natural Language Processing (NLP); Natural Language Interface (NLI); Entity Recognition (ER); Artificial Intelligence (AI); source code generation; pseudocode generation

## I. INTRODUCTION

Source-code is a list of human-readable instructions written in particular programming language. The aim of source-code is to check for precise specification, format and rules so that it can be interpreted into machine language [1]. Therefore, source-codes are the fundamentals of a computer program. It is usually written by a programmer or developer that has some training and knowledge of the programming language. There are many independent languages and each has its own distinctive way to writing instructions.

Natural Language Interface (NLI) provides a different input method in which users can interact with computer using spoken human language, like English instead of using a graphical user interface (GUI), command line interface (CLI) or computer languages like C and Python [2]. NLI enables the computer

to recognize and understand the flow of human language by providing an abstract layer that connects computers to users [3]. It enables users to enter their search queries in natural language which can be in either spoken audio or written text. The goal for most natural language systems is to make the system easier to use and to provide an interface that decrease the training time required for users.

The proposed system aims to generate source code of various programming languages like python and C using natural language as input. Making use of NLI to generate source-code can help beginners understand the language well. It can also help professionals to increase their working speed as uncommon and easy problems can be solved without going through documentation of that programming language.

As writing source-code is becoming more widespread and complicated, it is becoming an essential to automate writing simpler source-code by AIs to save time in writing, learning and understanding source-code. With Natural Language Processing (NLP) getting better and simple with each year and the demand for writing new and complex source code is increasing every year, it was inevitable for these two fields to join. There are many natural language interfaces that connect NLP with databases but not many programs that connect NLP with computer languages. Creating an Natural Language Interface that helps programmers to create source code will reduce the time it takes for them to write source code as they will refer to complex documentation less frequently and will reduce the bar to enter the world of programming language.

## II. LITERARY STUDY

### A. Different Natural Language Interface Approaches

Review of different approaches in natural language interfaces to databases [4] published by Reshma E. U. and Remya P. C. explores some Natural Language Interfaces to Databases (NLIDB) trends in 2017. They found that current NLIDB system consists of following types: 'Pattern matching' (i.e. Using manually defined rules), 'Syntax based system' (i.e. Creating parse tree and mapping it to database), 'Semantic grammar system' (i.e. Passing user input with hard wired semantic grammar and then creating parse tree which will be mapped to database) or 'Intermediate representation system' (i.e. it first translates the natural language input into intermediate logical query and then, it translates intermediate logical query into database query language.

There are many such systems that follow the same techniques and patterns with minor changes, namely, "IQS - Intelligent Querying System using Natural Language Processing" [5] and "MyNLIDB: A Natural Language Interface to Database" [6].

The advantages of these NLIDBs are that the users does not need to learn any artificial language, no need for spending time on training, simple and easy to use, are better for some questions and have high fault tolerance.

The disadvantages of these systems are that they deal with small amount of natural language i.e. can recognize limited set of words, errors and failures are not properly handled, ambiguity i.e. one word having many unrelated meaning can cause the query to change its meaning and finally users may not construct the query using recommended or pre-programmed words.

### B. Natural Language Query to SQL

In the paper "Formation of SQL from Natural Language Query using NLP" M. Uma et. al. [7] used the following techniques to extract information from natural language.

First, they extracted 'attribute' using Parts Of Speech (POS) tags. They used tokens next to a proper noun (i.e. NNP tag in NLTK Library). To search for 'date' they used regular expression (RegEx) to extract it using common writing formats. To extract 'fares' they used lemmatized word 'fare' and finally they used RegEx again for extracting train names.

This system is very rigid and can only perform SQL tasks on predefined database. But, we can use these methods to tag our own data to give to an AI model

### C. Conversion of Natural Language Query to SQL Query

In the paper titled "Conversion of Natural Language Query to SQL Query" by Abhilasha Kate et. al. [8] they first performed "Tokenization" on their input sentence and remove the stop words, then those tokens would be passed onto "Lexical analysis" which will replace all the words with their dictionary counterpart. This is the step where the natural language starts to look like a SQL sentence. Lastly, "Semantic Analysis" is performed which will replace natural sentence (e.g. less than or equal to) to their symbol counterparts (i.e.  $\leq$ ).

The given system has a shortcoming in lexical analysis phase as all the keywords needs to be known beforehand to be able to match those keywords with dictionary.

### D. Language to Code

The paper titled "Language to Code with Open Source Software" [9] published by Lei Tang, Xiaoguang Mao and Zhuo Zhang uses an encoder-decoder technique to automatically train NLP to generate source-code. They first converted the natural language descriptions into word-embeddings and fed it into the encoder to generate coding vector. Then the decoder maps this vector back into the desired code. They used LSTM neural network to train their model. Due to the labor-intensive nature of generating dataset they used a previously proposed method by Gu Xiaodong et.al in the paper "Deep code search" [10]. In this paper they proposed a unique method

to create training dataset i.e. they used comments from a Java program and its attached code snippet from open source Java projects as the dataset.

Even though this technique covers many complicated code scenarios, this technique is limited by the programming language it generates (here they can only generate Java source-code). To generate source-code in other programming languages, we need to create another database from scratch which is a lengthy and tedious task. From this system we can adopt the training methods they used with the databases tagging techniques they used.

### E. Natural Language Database Query Interface

"A Simple Guide to Implement Data Retrieval through Natural Language Database Query Interface (NLDQ)" [11] published by Tameem Ahmad and Nesar Ahmad uses a straight forward approach to convert a natural language statement to database query. They first used "Tokenizer" to divide the input into individual tokens or words. They then used "Parsing" to create a parse tree with its related POS (parts-Of-Speech) tags. After this they used "Syntactic Comparison" to check if any keywords that appeared in the input is already available in the database as either a direct match or a alias of it. Lastly, in "SQL Generator" they used static templates that can be used as fill-in-the-blanks to choose the correct template and fill all the related fields labeled in previous steps.

The main advantage of this type of system is that it is easy to make any new changes that the client requests. But the downsides are that this is a very rigid system that is tied to a particular database. Although, we won't tie our system with templates, but each programming language uses a predefined format (i.e. main function, indentation, parentheses "}", etc.) which we can add to our system.

### F. Modified Co-occurrence Matrix Technique

Anuradha Mohite and Varunakshi Bhojane [12] proposed a updated way to find co-occurrence matrix formulation in the paper titled "Natural Language Interface to Database Using Modified Co-occurrence Matrix Technique". They first parsed the input to create POS (i.e. Parts Of Speech) tags and parse tree. They then used modified Hyperspace Analogue to Language (HAL) matrix to find tokens related to nouns. Using cosine-similarity they get a table with nouns associated with different POS and using this technique they identified the WHERE clause in SQL (Structured Query Language).

With the hardest part of the query now identified they used stemming technique to convert all words into their common roots and then used semantic mapping to find words such as min, max, avg,  $\geq$ , etc. Once these word are identified they used bi-gram algorithm to find correct attribute and table name from the database and create the final query to be displayed to user along with its output. "Natural Language to Structured Query Language using Elasticsearch for descriptive columns" [13] uses similar approach with changes made in word embeddings.

Their clever use of POS tag pair such as "numeric value-noun pair" or "proper noun-noun pair" to find where clause in SQL can help us to identify inconsistent attributes such as names of variables and functions in our system.

### G. Pseudocode to Source-code

Teduh Dirgahayu et.al. [14] proposed a method to automatically convert pseudocode to Source-code. In this method, the pseudocode is first translated to an intermediate model then to source-code. The intermediate model consists of a parse tree that is created with the help of a tool called ANTLR. This represents pseudocode in a more structured and language independent way. Then language dependent tool for generating source-code is created.

The main shortcoming of this method is that the pseudocode which is in the form of XML needs to be created manually. This manual pseudocode and source-code must comply to their respective grammars (i.e. metamodels). The XML intermediate model must comply to a XML schema.

### H. An XML-based Pseudo-code Online Editing and Conversion System

Liu Haowen et.al in their paper "An XML-based Pseudo-code Online Editing and Conversion System" [15] introduced an innovative way to convert pseudo code into source-code. They first convert the input pseudo code into an XML (eXtensive Markup Language) format using DOM4J package available in Java programming language. From there, using the same DOM4J package, they converted the pseudo code in XML format into Java source-code. In doing so, they also created XML tags for pseudo code which we can use in our system as a reference.

The reason to choose XML is that it is a cross-platform language i.e. they can be used in any software and in any operating system making this system independent of any programming language.

### I. Pseudocode to Source-code using NLP

Ayad Tareq Imam et.al [16] proposed a unique way to convert pseudocode to source-code i.e. using NLP. The problem with this method is that it presents a complex solution to a simple problem. As seen in previous 2 papers [14] & [15], the same output can be achieved with simple one-to-one mapping.

### J. Generating Source-code without NLP

Till here we have studied ways to generate structured code like SQL queries (due to lack of source-code generation) using NLP techniques. There are other papers that can generate proper source-code but they bypass the NLP requirements and directly train their models to generate source-code from training data.

Some such papers are "DeepCoder" [17], "Text2App" [18], "Generating Pseudo-Code from Source Code Using Deep Learning" [19], "Incorporating External Knowledge through Pre-training for Natural Language to Code Generation" [20], "In-IDE Code Generation from Natural Language" [21] and "Programming with a Differentiable Forth Interpreter" [22].

All these papers faces a problem of lack of source-code datasets. Some uses premade dataset, many generate their own. Another problem they face is all these papers generate source-code in one specific programming language

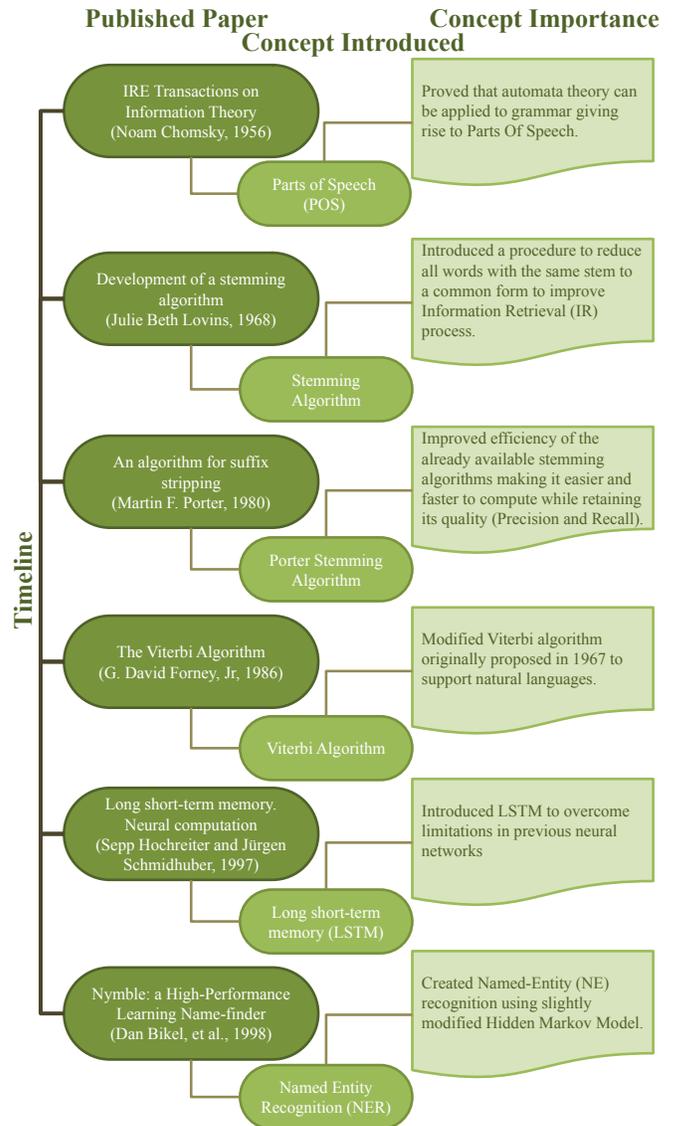


Fig. 1. History of NLP Concepts.

### K. History of Various Concepts of NLP

Fig. 1 lists the papers where important concepts have been introduced in history of NLP with the importance of the paper. These concepts have been used in our system. The Long Short-Term Memory (LSTM) neural network [23] has not been mentioned explicitly in our system but we used this neural network to train our various NLP models.

Some may argue that the NLP as a field began when Warren Weaver mentioned using of modern computing devices to translate from one language to another in his memorandum [24] in 1949 and the rest is history.

1) *Parts of Speech*: Noam Chomsky in 1956 [25] introduced the concept that grammar of a language can be viewed as a theory of the structure of this language and is based on certain finite set of observations. It introduced the finite-state language for NLP that we called today as Parts Of Speech (POS).

2) *Stemming Algorithm*: The stemming algorithm which converts all the redundant words to its common "stem" was published in 1968 by Julie Beth Lovins [26]. Surprisingly its main purpose was not for NLP purposes but for retrieving huge amount of data from databases (back then data transfers were a lot slower).

3) *Porter-Stemming Algorithm*: The famous porter-stemming algorithm was easy to find as it is in public repository online. This was developed to further improve the efficiency of the information retrieval systems of the various stemming algorithms available at that time [27].

4) *The Viterbi Algorithm*: Although, Viterbi algorithm was originally proposed to calculate the error bounds in convolutional codes [28] as just a proof of concept, David Forney Jr. [29] in his paper "The Viterbi Algorithm" modified this algorithm for NLP purposes. Till this date we use this algorithm alongside with Hidden Markov Model to predict the POS tag of a word with probability even though an AI solution exists.

5) *Named Entity Recognition (NER)*: "Nymble: a High-Performance Learning Name-finder" published by Daniel Bikel et.al. [30] was the first to introduce a Named-Entity Recognition system using slightly modified Hidden Markov Model. It performs at or above the 90% accuracy level, often considered "near-human performance".

Observing all those systems, we came to conclusion that there is a lack of NLP system where a natural language is converted into a source-code (all those systems converted natural language into database queries). Also, those systems that do convert to natural language does not allow customizations (i.e. adding custom keywords to recognize) as a feature.

Our project intends to introduce a novel method that helps the programmers in developing source-code and increasing their speed and efficiency.

### III. PROBLEM STATEMENT

The project intends to introduce a novel method that helps the programmers in developing source-code increasing their speed and efficiency.

- An interface to take natural language as input.
- An option to select programming language on the interface (initially python).
- Convert natural language into pseudo code using NLP (using intent recognition and entity recognition).
- Convert pseudo code into source-code (using traditional programming).
- Output the source-code onto the interface.

### IV. PROBLEM DEFINITION

For a programmer, the ability to learn a programming language's commands and functions on the fly is crucial to the time they spend on creating a source-code. No matter the skill of the programmer, there are always cases where they have to tackle uncommon features present in a programming language which they have not seen before. To solve those uncommon and unseen features they refer documentation of

that programming language and then learn from it which requires time.

The programmers can define the problem in natural languages easily, the problem lies in remembering the exact keywords and parameters. By providing a Natural Language Interface we can shorten the time the programmer takes to search through documentations. It can also help to reduce the skill and experience required to write complicated source-code which heavily depends on using functions.

### V. PROPOSED WORK

- 1) A python based interface to input English natural language and an option to choose target programming language.
- 2) Gather training data from the internet. The data will be simple beginner's problem statements from various sites for variety.
- 3) Provide "Parts Of Speech" (POS) tags to each word.
- 4) Train "Intent Classification" to recognize different operations such as "addition", "multiplication", "variable declaration", "print statement", etc.
- 5) Train "Entity Recognition" to recognize various data types of variables.
- 6) Identify variable names through POS if they are present in the input.
- 7) Write a function for each "intent" and extract information like number of variables, their type and name if it is present and other information if required.
- 8) Convert the extracted information into a pseudo code in a XML format.
- 9) Read that XML file and convert it into source-code through python.
- 10) Save the source-code into a file.
- 11) Display the file into the interface.

### VI. LIBRARIES REQUIRED

#### A. Tkinter

Tkinter is the standard GUI library for Python. Tkinter provides a fast and easy way to create GUI applications [31]. Tkinter provides various controls, such as buttons, labels and text boxes used in a GUI application. These controls are commonly called widgets.

#### B. NLTK vs SpaCy

NLTK (Natural Language Toolkit) [32], spaCy [33] and Stanford's CoreNLP [34] are all similar libraries that provide off-the-shelf functions for NLP. As we are using "python" programming language for creating a demo, we need to list all pros and cons on NLTK and spaCy libraries as those support the python language, coreNLP does not (it only supports Java). Table I lists the differences between NLTK and spaCy libraries.

The paper titled "Using Natural Language Processing to Detect Privacy Violations in Online Contracts" by P. Silva et.al. [35] and another paper titled "Extractive Automatic Text Summarization using SpaCy in Python & NLP" [36] made comparison between spaCy, coreNLP and NLTK and came to conclusion that coreNLP has the best performance in terms of precision, recall and F1 score followed by spaCy and lastly NLTK.

TABLE I. NLTK LIBRARY VS SPACY LIBRARY

| NLTK                                                                                                                  | SpaCy                                                                                                                                                                             |
|-----------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| NLTK (Natural Language Toolkit) library was released in 2001.                                                         | SpaCy library was released in February 2015.                                                                                                                                      |
| NLTK library is only supported in "python" programming language.                                                      | SpaCy is supported in number of programming languages like "R", "ruby", "cpp", "java script", ".net", "python", etc.                                                              |
| NLTK has a lot of algorithms as compared to spaCy which is helpful in learning and research applications.             | SpaCy has a small group of algorithms that they regularly update, which is critical in industry usage.                                                                            |
| As it is not updated regularly, the algorithms provided by NLTK are comparatively slower than spaCy.                  | With their regular update to latest techniques spaCy is significantly faster than NLTK.                                                                                           |
| NLTK incorporates several languages.                                                                                  | SpaCy have statistical models for seven languages including English, German, Spanish, French, Portuguese, Italian, and Dutch, It also braces "named entities" for multi-language. |
| NLTK is string processing library. It takes input as strings and provides output as string or lists of strings.       | SpaCy uses object-oriented approach. It takes input in string but return the output in objects.                                                                                   |
| NLTK does not support word vectors.                                                                                   | SpaCy has support for word vectors.                                                                                                                                               |
| NLTK tries to break the text into sentences. It just returns the words itself, no extra information is given with it. | SpaCy builds a semantic tree for individual sentence as higher a potent approach, returns more information.                                                                       |
| NLTK has inferior precision, recall and F1 score than spaCy.                                                          | SpaCy has better precision, recall and F1 score than NLTK.                                                                                                                        |

### C. Regex

Although spaCy provides "Pattern Matcher" functionality to its toolbox which is similar to regex, a regular expression or regex can be used in some limitations of spaCy's pattern matcher.

A regex defines a set of strings that lets you check if a particular string is present in a large text [37]. The most common usage of regex is to alert system administrators if an error appeared in log files. Another example would be to extract phone numbers, email addresses from a large database. It is commonly used where fixed pattern appears.

To find a email address that has the format "someone@somedomain.sometopleveldomain" (eg. john@gmail.com) we use the regex "[a-zA-Z0-9+.-]+@[a-zA-Z0-9-]+.[a-zA-Z0-9-]+". To find emails in plethora of texts.

## VII. WORKING MODEL

The main goal of our system is to convert "Natural Language" statements into source-code. It is able to handle one line statements as input and can generate 3-5 lines source-code depending on the input given.

Fig. 2 shows the working of our system when the system has been deployed at the client side. It is a step-wise working model which defines the processes from user input to showing output to the user and all the other steps in-between. In this figure, the light-blue nodes denotes the generation of files. For example, the node "Pseudo code statement/command" denotes saving the .xml file in log folder. The white nodes denotes processing.

### A. NLP Translator

The first step in this process is to get the user input. The input has to be in English natural language and has to describe

the programming problem. Once the input is received the "NLP Translator" uses a trained NLP model to break down the input and extract important information from it. In this process, the first thing applied is "Tokenizer" to split the input into individual tokens or words, then entity extractor identifies important keywords like equals sign, condition statements, etc. Next, the "Intent Identifier" identifies what kind of operation does the user wants to perform like "addition", branching, etc. Lastly, POS is used to identify the variable name, its value (if present) and function/program name.

### B. Post Processor

Once all the input has been identified, it's time to convert the extracted information into a pseudo code, this is done is "Post Processing" step. It first arranges the input in specific order that resembles a pseudo code, adds extra information if missing and the using XML parsers converts the pseudo code into XML. After this we are left with pseudo code file that is easier to read by both humans and computers thanks to the XML format.

### C. Rule Based Translator and Its Post Processor

Rule Based Translator is an easier step that reads the pseudo code line by line and converts the keywords from universally understood to language specific format. The post processor adds extra features to the syntactical code like parenthesis "{ }", indentation, semi-comma ";", etc. The output is then saved on disk and displayed to the user on the interface.

## VIII. ALGORITHM

Algorithm given in Fig. 3 is a coarse-grain view of the proposed system that gives the basic steps needed to implement this system.

We first start by creating an interface that can get input and provide output. Second step is to search for a dataset. Since, we did not find any we created the dataset from scratch. The method of which is defined further ahead.

Next is to create a Named Entity Recognition or NER model. To create a NER model we first need to provide Parts Of Speech or POS tags which then be used by NER model to identify various entities.

Once the system is trained we then start extracting information from it. The first information we extract is variables, their name, datatype, scope and value. After which "operations" are identified and all the variables associated with it are processed i.e. each individual part of the operation is identified (for example, in the statement "add var1 and var2", add is the type of operation, var1 and var2 are the two variables associated with it).

The extracted information is then parsed into XML format and a pseudocode file is generated. Then this XML pseudocode file is read again by the system and is converted into source-code. This source-code is saved on the user's system and displayed on the interface.

## IX. SYSTEM DESIGN

Fig. 4 is a detailed view of the proposed system that explains the steps needed to create this system.

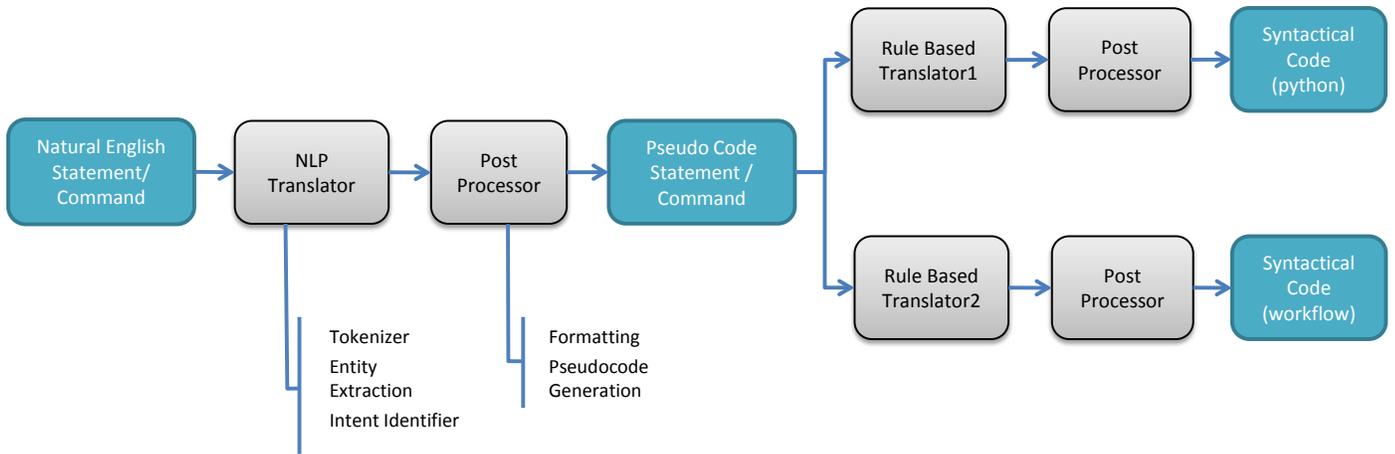


Fig. 2. NLI-GSC Working Model.

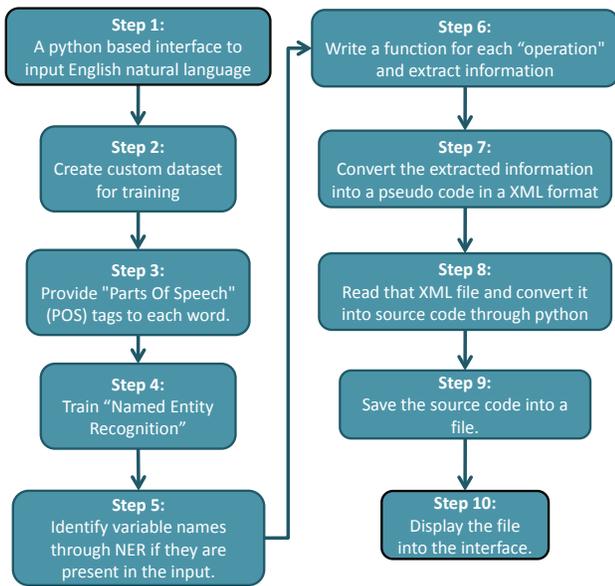


Fig. 3. Algorithm of NLI-GSC System.

A. Data Generation

There are not many NLP datasets, if any, that address the programming aspects that we can use. Generating data manually for an AI system is very tedious and time-consuming process as it requires a minimum of hundreds of data to provide any acceptable results. Hence, due to these reasons we have developed an automated system that can create its own dataset.

As seen in Fig. 4 & 5, we first define "nutrients" tokens i.e. variations of similar word. For example, "DATATYPE" is a nutrient containing integer, int, float, str, string, etc. We then define "seed" sentences. Seed sentences are made up of combinations of "nutrients". An example of seed sentence is "COMMAND DATATYPE VARIABLE" which when expanded will result in sentences like "create integer var1", "define float area" and "initialize string text\_1".

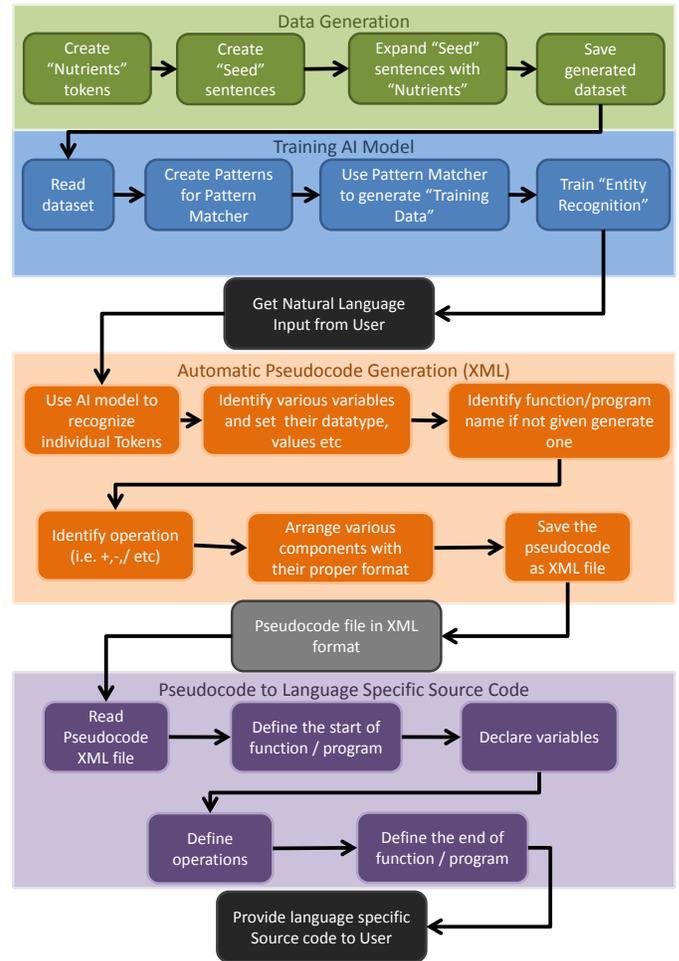


Fig. 4. NLI-GSC System Design.

The next step is to expand the seeds using individual nutrient tokens. This will result in generating all the possible combination of statements for the given format. In our program, we were able to generate 1,565,668 statements from 23 seed statements (each seed's length ranging from 3-6 nutrient

tokens) and 19 individual nutrient tokens. Since, 1.5 mil is too much of a dataset, we have reduced it with the ratio 1:20 resulting in getting a final dataset of 78,284 statements.

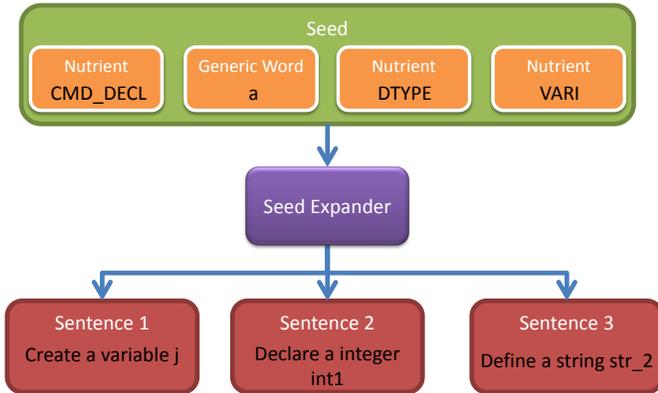


Fig. 5. Data Generation Process.

### B. Training AI Model

For training of the AI model, we used "spaCy" library [33]. Unlike its NLTK counterpart which focuses on academics and learning, spaCy is built for industrial uses and production.

We used Named Entity Recognition (NER) as our model to tag words that may belong to certain instruction. For example, the word "int" will tell NER model that it belongs to "datatype" while words like "addition", "sub", "\*" will be recognized as "operations".

In training in the AI model, we first need to create a dataset from data which we have achieved using spaCy's "Pattern Matcher". Since, we can control which words appear in the data in Data Generation phase, we can easily use pattern matcher to find a word or sequence of words to tag it. Once all the relevant words are tagged and a training dataset is created we can proceed to train it. Again, we used spaCy standard libraries to train. We used 3 epochs with batch size = 100 and got the final training loss of 1.95.

### C. Automatic Pseudocode Generation

When the user inputs their natural language query, our system automatically converts it to pseudocode and saves it in XML format. The process is as follows.

After the user has given their input, the system uses NER model trained in previous step to identify various tags examples can be seen in Fig. 6 and 7. Once the tokens have been recognized, we lemmatize them for easy recognition. We lemmatize all the types of tokens inside the tags except variables tag, since they are user given names. We process them based on the tags it has been assigned.

The variables are assigned names (if none is given it will automatically generate based on datatype), datatypes (If none is given, it will first try to get it from another variable in the input if not it will be assigned as string.), scope (local or global) and values (if they are given). If the program recognizes that it does not require any variables it will skip this step.

The input is recognized as either a function or a program and then the name of the function/program is recognized if the user has mentioned it in the query. A name is automatically generated based on available tags if the user has not mentioned the name explicitly. Once an operation tag is detected it will take all the variables associated and arrange them in specific pattern (i.e.  $ans = var1 + var2$ );).

Finally, all the components will be arranged by the order of appearance. First, the program/function will be declared with a proper indentation, then variables will be declared and then operations will be written and lastly if the language demands, the program/function will end (For example, some program requires "}" to end them ). All this will then be parsed to XML file and saved on the desired location.

### D. Pseudocode to Language Specific Source-code

Once the pseudocode file is ready, the only thing left to do is to convert it into a programming language. Using techniques that convert XML to program [14] [15], we can easily convert pseudocode to source-code. An automatic pseudocode generation is independent in creating a language specific source-code. Hence, we can- create the output of any programming language by creating a new file.

It's a simple process that reads the XML from top to bottom and depending on the tag encountered it will use a specific format to fill in the blanks.

## X. RESULTS

Our system produces multiple outputs before reaching the final source-code output. Below are the outputs according to the order they are generated.

### A. Generating Data for Training

As explained in Section IX-A, seed statements are used to generate input data that is later used to train the NLP model. Table II lists all the seed statements used in the system along with some sample output it generates. This output is unlabelled text data that acts as raw dataset.

TABLE II: Seed Statements along with Example Uoutput

| Seed No. | Seed Statement                                                               | Example Output                                        |
|----------|------------------------------------------------------------------------------|-------------------------------------------------------|
| 1        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE                          | Write a function to creates global integer            |
| 2        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_NUM equal to<br>RAND_NUM | Create a program to define global integer equal to 20 |

Continued on next page

TABLE II: Seed Statements along with Example Uutput (Continued)

| Seed No. | Seed Statement                                                                        | Example Output                                                   |
|----------|---------------------------------------------------------------------------------------|------------------------------------------------------------------|
| 3        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_NUM<br>VARI=RAND_NUM              | Write a program for create number k=18                           |
| 4        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE VARI                              | Create a function to declares integer into                       |
| 5        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_NUM to<br>RAND_NUM                | Create program for create local nums to 59                       |
| 6        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE equal VARI to<br>RAND_NUM         | set local variable equal j to 77                                 |
| 7        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_CHAR = 'hi'                       | Write function to set characters = 'hi'                          |
| 8        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_CHAR to 'hi'                      | set characters to 'hi'                                           |
| 9        | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR VARI = 'hi'                                | Write a function to creates global lists1 = 'hi'                 |
| 10       | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR VARI to 'hi'                               | Write a function for create global txt_1 to 'hi'                 |
| 11       | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_CHAR VARI = 'please enter value'  | Write a program for set local char char_0 = 'please enter value' |
| 12       | FUNC_PROG_START<br>CMD_DECL<br>SCOPE_OFVAR<br>DTYPE_CHAR VARI to 'please enter value' | create character j to 'please enter value'                       |
| 13       | FUNC_PROG_START<br>CMD_PRNT VARI                                                      | Write function to show strso                                     |
| 14       | FUNC_PROG_START<br>CMD_PRNT 'please select value'                                     | show 'please select value'                                       |

Continued on next page

TABLE II: Seed Statements along with Example Uutput (Continued)

| Seed No. | Seed Statement                                                                         | Example Output                                                |
|----------|----------------------------------------------------------------------------------------|---------------------------------------------------------------|
| 15       | FUNC_PROG_START<br>CMD_PRNT 'please select value' + VARI                               | write 'please select value' + str1                            |
| 16       | FUNC_PROG_START<br>CMD_PRNT VARI + i                                                   | display c + i                                                 |
| 17       | FUNC_PROG_START<br>CMD_PRNT 'please select value' + i                                  | Create a program to shows 'please select value' + i           |
| 18       | FUNC_PROG_START<br>CMD_INPUT VARI                                                      | Create function for input string_1                            |
| 19       | FUNC_PROG_START<br>CMD_INPUT 'please enter name'                                       | Create function for insert 'please enter name'                |
| 20       | FUNC_PROG_START<br>add VARI + VARI                                                     | Create a function to add strings0 + strings_0                 |
| 21       | SPL_FUNC VARI to SPL_CLASS                                                             | print strso to screen                                         |
| 22       | Write a FUNC_PROG<br>FUNC_NAME to perform OPER with VARI, DTYPE_NUM<br>VARI = RAND_NUM | Write a program prog_div to perform add with i, floats j = 47 |
| 23       | FUNC_PROG_START<br>OPER SCOPE_OFVAR<br>DTYPE_NUM VARI<br>and VARI                      | Create a function for mod local number c and address          |

Table III depicts time taken for each seed statement to generate their respective statements, their individual time and the amount of raw statements it generates. The full seed statements can be seen in Table II. It can observe that statements that have large number of nutrients takes more time to execute and generate more statements compared to ones that have small number of nutrients.

### B. Varying NLP Hyper-parameters

While training this NLP models, there were various hyper-parameters like learning rate, batch size, dropout rate, total epochs, etc. Changing those hyper-parameters resulted in different loss and execution time. In Table IV, shows various effects of changing those hyper-parameters.

The first column is kept as the base for benchmark as this is giving the best possible loss value along with relatively fast training time. The other columns shows gradual changes in various parameters, the changed parameters are shown as **bold** while the unchanged parameters (compared to first column) are normal.

TABLE III. TIME TAKEN TO GENERATE INPUT DATA STATEMENTS FROM SEED STATEMENT ALONG WITH THE AMOUNT EACH STATEMENT GENERATES

| Seed No. | Individual Time | Cumulative Time | No. of Statements Generated |
|----------|-----------------|-----------------|-----------------------------|
| 1        | 0.049s          | 0.049s          | 15288                       |
| 2        | 0.022s          | 0.073s          | 7176                        |
| 3        | 0.438s          | 0.512s          | 174096                      |
| 4        | 0.383s          | 0.900s          | 271440                      |
| 5        | 0.031s          | 0.931s          | 7176                        |
| 6        | 0.721s          | 1.653s          | 271440                      |
| 7        | 0.014s          | 1.668s          | 4056                        |
| 8        | 0.015s          | 1.684s          | 4056                        |
| 9        | 0.077s          | 1.762s          | 91728                       |
| 10       | 0.078s          | 1.840s          | 91728                       |
| 11       | 0.150s          | 2.006s          | 136656                      |
| 12       | 0.172s          | 2.179s          | 136656                      |
| 13       | 0.015s          | 2.195s          | 30576                       |
| 14       | 0.014s          | 2.210s          | 104                         |
| 15       | 0.031s          | 2.241s          | 30576                       |
| 16       | 0.027s          | 2.269s          | 30576                       |
| 17       | 0.000s          | 2.270s          | 104                         |
| 18       | 0.009s          | 2.279s          | 19110                       |
| 19       | 0.000s          | 2.279s          | 65                          |
| 20       | 0.000s          | 2.279s          | 3822                        |
| 21       | 0.000s          | 2.279s          | 1470                        |
| 22       | 18.621s         | 20.901s         | 8195904                     |
| 23       | 0.552s          | 21.700s         | 369954                      |

TABLE IV. VARIOUS HYPER-PARAMETERS CONFIGURATION AND THEIR EFFECTS BASED ON EPOCHS

|                  | Learn Rate: 0.001                   | Learn Rate: 0.001                  | Learn Rate: 0.001                 | Learn Rate: 0.005                | Learn Rate: 0.005                  | Learn Rate: 0.001                  |
|------------------|-------------------------------------|------------------------------------|-----------------------------------|----------------------------------|------------------------------------|------------------------------------|
|                  | Batch Size: 1000                    | Batch Size: 1000                   | Batch Size: 5000                  | Batch Size: 1000                 | Batch Size: 5000                   | Batch Size: 5000                   |
|                  | Dropout Rate: 0.0                   | Dropout Rate: 0.1                  | Dropout Rate: 0.0                 | Dropout Rate: 0.0                | Dropout Rate: 0.1                  | Dropout Rate: 0.1                  |
| <b>Epoch 1/5</b> | Losses: 143006.98<br>Time: 90.46s   | Losses: 173171.71<br>Time: 102.71s | Losses: 576695.25<br>Time: 85.97s | Losses: 52150.45<br>Time: 91.47s | Losses: 278816.12<br>Time: 101.01s | Losses: 611403.16<br>Time: 99.14s  |
| <b>Epoch 2/5</b> | Losses: 14.13<br>Time: 178.95s      | Losses: 14.34<br>Time: 206.14s     | Losses: 89620.88<br>Time: 172.35s | Losses: 95.10<br>Time: 188.07s   | Losses: 435.87<br>Time: 201.79s    | Losses: 200406.14<br>Time: 199.16s |
| <b>Epoch 3/5</b> | Losses: 7.49<br>Time: 269.70s       | Losses: 0.94<br>Time: 313.47s      | Losses: 1655.89<br>Time: 259.04s  | Losses: 219.68<br>Time: 298.05s  | Losses: 81.57<br>Time: 306.28s     | Losses: 9531.23<br>Time: 300.40s   |
| <b>Epoch 4/5</b> | Losses: 0.00014<br>Time: 363.22s    | Losses: 2.53<br>Time: 422.21s      | Losses: 11.83<br>Time: 348.93s    | Losses: 54.44<br>Time: 424.99s   | Losses: 87.22<br>Time: 414.40s     | Losses: 481.82<br>Time: 405.47s    |
| <b>Epoch 5/5</b> | Losses: 1.6228e-06<br>Time: 457.60s | Losses: 0.0014<br>Time: 531.43s    | Losses: 1.77<br>Time: 440.41s     | Losses: 2.00<br>Time: 563.29s    | Losses: 26.73<br>Time: 524.79s     | Losses: 7.43<br>Time: 511.42s      |

C. Input Statement

We used the below 2 statements to demonstrate the output of our system during various stages of the system. These are the example of inputs that the user might use during its production phase.

Statement 1:

define variable text1 = 'please enter value'

Statement 2:

create a function to add a and local number num\_1 = 15

and b

D. NER System Output

Fig. 6 and 7 are the outputs given by the NER model. Each word (token) is bundled with original word given by the user and tag assigned by the NER model. In the left is the word/token and in the right is tag assigned.

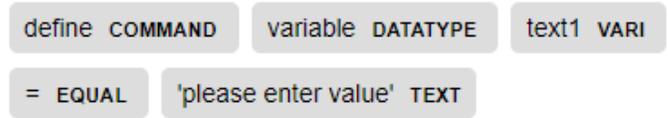


Fig. 6. NER Model Output of Statement 1.



Fig. 7. NER Model Output of Statement 2.

E. Pseudocode Output

Pseudocode file generated by our system is shown below in listings 1 and 2.

```

1 <?xml version="1.0" ?>
2 <root>
3 <statement value="define variable text1 = 'please
4 enter value'"/>
5 <program type="program" name="define_variable">
6 <variables>
7 <variable name="text1" datatype="variable"
8 value="'please enter value'"/>
9 </variables>
10 </program>
11 </root>

```

Listing 1: Pseudocode output of statement 1

```

1 <?xml version="1.0" ?>
2 <root>
3 <statement value="create a function to add a and
4 local number
5 num_1 = 15 and b"/>
6 <program type="function" name="create_function">
7 <variables>
8 <variable name="a" datatype="number"/>
9 <variable name="num_1" scope="local"
10 datatype="number" value="15"/>
11 <variable name="b" datatype="number"/>
12 </variables>
13 <add variable1="a" variable2="num_1" variable3
14 = "b"/>
15 </program>
16 </root>

```

Listing 2: Pseudocode output of statement 2

F. Final Source-code

The source-code is the final output generated. This is the only output visible to the user. Listings 3 and 4 shows the output code.

```
1 def main():
2 variable text1 = 'please enter value'
3
4 if __name__ == "__main__":
5 main()
```

Listing 3: Source-code output of statement 1

```
1 def create_function():
2 number a
3 number num_1 = 15
4 number b
5
6 answer = a + num_1 + b
7
8 create_function()
```

Listing 4: Source-code output of statement 2

The final results as obtained in our program that is made using python can be seen in Fig. 8, 9, 10 and 11.

## XI. FUTURE WORK

One of the ways we can further improve on this project is to create support for more programming languages. Currently we have only implemented support for python and C. We can also add support for new common functions like date, time, string operations, etc. Loops like "for", "while" and "do while" are also left out due to its complexity and time constraints which can be expanded later.

Our approach serves as the basic idea that allows the development of systems that are more complex in terms of keyword detection and contains more functionalities like loops and branching by adding additional NLP elements like dependency parsing and Semantic parsing.

## XII. CONCLUSION

This proposal shows that a language-independent solution is a feasible alternate for writing source-code without having full knowledge about a programming language.

Using XML based pseudo code as an intermediate step makes this method as programming language-independent which solves the major drawbacks in existing research that comes with a rigid commitment to only one programming language. The next step, which is converting XML based pseudo code into language-dependent source-code is dependent on premade language format which is modifiable by anyone, making this approach module based approach. If a person wishes to convert the natural language into some other programming language, they simply need to duplicate the premade language format and fill it with their desired programming language keywords.

Another challenge faced in this area which is program based NLP is that, there is not many datasets available in this particular sub-field which severely limits the research capabilities and keeps this sub-field from growing forward. Although not ideal, our dataset generation system provides an automated approach to create thousands of data that can be used in an AI system as a training dataset.

One way this system can be used is with ticket based programming, where the programmers get their tasks in the

form of tickets in their mail. The system can be used as a suggestion system where the mail is analyzed and a suggested solution is provided to the programmer. Another use is to pair this system with voice recognition and let the programmer write simple source-code only through speech making their hands free for writing other more complex code.

## REFERENCES

- [1] S. Oualine, *Practical c Programming, Third Edition*, 3rd ed. O'Reilly Media, Inc., 1997.
- [2] C. W. Thompson and K. M. Ross, "Natural-language interface generating system," in U. S. Patent. U.S., 1987.
- [3] S.-Y. Park, J. Byun, H.-C. Rim, D.-G. Lee, and H. Lim, "Natural language-based user interface for mobile devices with limited resources," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2086–2092, 2010.
- [4] E. U. Reshma and P. C. Remya, "A review of different approaches in natural language interfaces to databases," in *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, 2017, pp. 801–804.
- [5] P. Gupta, A. Goswami, S. Koul, and K. Sartape, "Iqs-intelligent querying system using natural language processing," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 2, 2017, pp. 410–413.
- [6] A. Das and R. C. Balabantaray, "Mynlidb: A natural language interface to database," in *2019 International Conference on Information Technology (ICIT)*, 2019, pp. 234–238.
- [7] M. Uma, V. Sneha, G. Sneha, J. Bhuvana, and B. Bharathi, "Formation of sql from natural language query using nlp," in *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019, pp. 1–5.
- [8] A. Kate, S. Kamble, A. Bodkhe, and M. Joshi, "Conversion of natural language query to sql query," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 2018, pp. 488–491.
- [9] L. Tang, X. Mao, and Z. Zhang, "Language to code with open source software," in *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSSESS)*, 2019, pp. 561–564.
- [10] X. Gu, H. Zhang, and S. Kim, "Deep code search," in *2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*, 2018, pp. 933–944.
- [11] T. Ahmad and N. Ahmad, "A simple guide to implement data retrieval through natural language database query interface (nldq)," in *2019 8th International Conference System Modeling and Advancement in Research Trends (SMART)*, 2019, pp. 37–41.
- [12] A. Mohite and V. Bhojane, "Natural language interface to database using modified co-occurrence matrix technique," *2015 International Conference on Pervasive Computing (ICPC)*, pp. 1–4, 2015.
- [13] S. S. Badhya, A. Prasad, S. Rohan, Y. S. Yashwanth, N. Deepamala, and G. Shobha, "Natural language to structured query language using elasticsearch for descriptive columns," in *2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, vol. 4, 2019, pp. 1–5.
- [14] T. Dirgahayu, S. N. Huda, Z. Zulkhri, and C. I. Ratnasari, "Automatic translation from pseudocode to source code: A conceptual-metamodel approach," in *2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 2017, pp. 122–128.
- [15] L. Haowen, L. Wei, and L. Yin, "An xml-based pseudo-code online editing and conversion system," in *IEEE Conference Anthology*, 2013, pp. 1–5.
- [16] A. T. Imam and A. J. Alnsour, "The use of natural language processing approach for converting pseudo code to c# code," *Journal of Intelligent Systems*, vol. 29, no. 1, pp. 1388–1407, 2020. [Online]. Available: <https://doi.org/10.1515/jisys-2018-0291>
- [17] M. Balog, A. Gaunt, M. Brockschmidt, S. Nowozin, and D. Tarlow, "Deepcoder: Learning to write programs," 11 2016.
- [18] M. Hasan, K. S. Mehrab, W. U. Ahmad, and R. Shahriyar, "Text2app: A framework for creating android apps from text descriptions," *ArXiv*, vol. abs/2104.08301, 2021.
- [19] A. Alhefdhi, H. K. Dam, H. Hata, and A. Ghose, "Generating pseudo-code from source code using deep learning," in *2018 25th Australasian Software Engineering Conference (ASWEC)*, 2018, pp. 21–25.
- [20] F. F. Xu, Z. Jiang, P. Yin, B. Vasilescu, and G. Neubig, "Incorporating

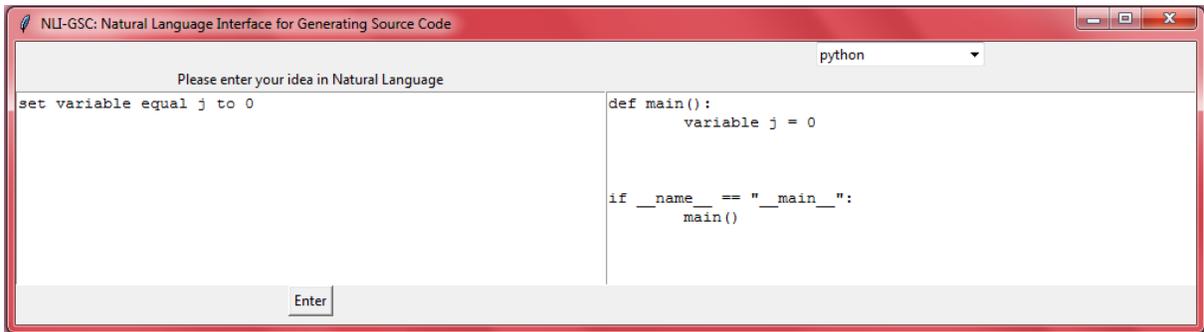


Fig. 8. Screenshot of our Program 1.

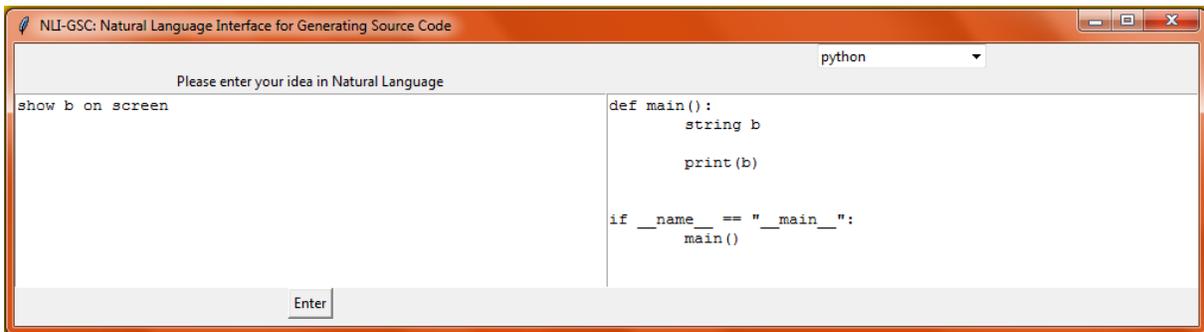


Fig. 9. Screenshot of our Program 2.

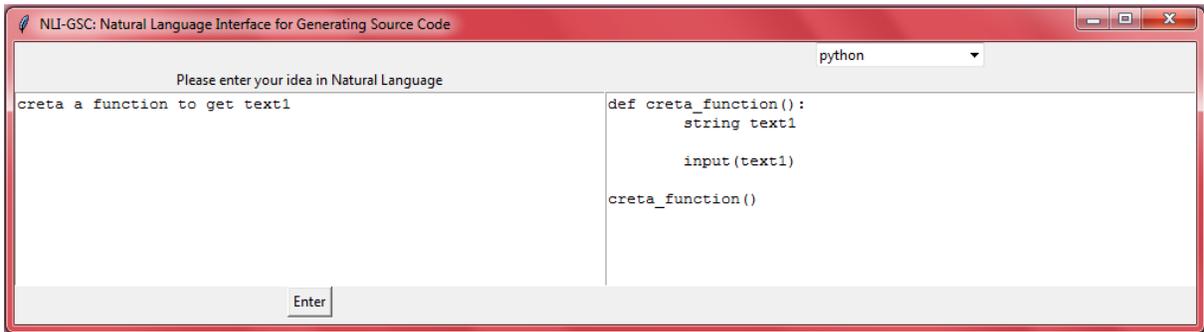


Fig. 10. Screenshot of our Program 3.

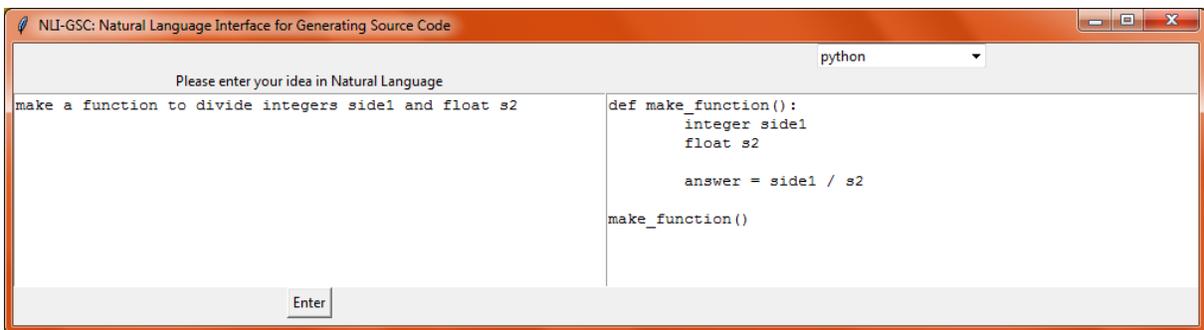


Fig. 11. Screenshot of our Program 4.

external knowledge through pre-training for natural language to code generation," in *ACL*, 2020.

[21] F. F. Xu, B. Vasilescu, and G. Neubig, "In-ide code generation from natural language: Promise and challenges," *ArXiv*, vol. abs/2101.11149,

2021.

[22] M. Bosnjak, T. Rocktäschel, J. Naradowsky, and S. Riedel, "Programming with a differentiable forth interpreter," in *ICML*, 2017, pp. 547–556.

- [Online]. Available: <http://proceedings.mlr.press/v70/bosnjak17a.html>
- [23] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.
- [24] W. Weaver, "Translation," in *Machine Translation of Languages*, W. N. Locke and A. D. Boothe, Eds. Cambridge, MA: MIT Press, 1949/1955, pp. 15–23, reprinted from a memorandum written by Weaver in 1949.
- [25] N. Chomsky, "Three models for the description of language," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113–124, 1956.
- [26] J. B. Lovins, "Development of a stemming algorithm," *Mech. Transl. Comput. Linguistics*, vol. 11, pp. 22–31, 1968.
- [27] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 40, pp. 211–218, 1980.
- [28] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.
- [29] G. Forney, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.
- [30] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learning name-finder," 04 1998.
- [31] F. Lundh, "An introduction to tkinter," URL: [www.pythonware.com/library/tkinter/introduction/index.htm](http://www.pythonware.com/library/tkinter/introduction/index.htm), 1999.
- [32] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.
- [33] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.
- [34] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60. [Online]. Available: <http://www.aclweb.org/anthology/P/P14/P14-5010>
- [35] P. Silva, C. Gonçalves, C. Godinho, N. Antunes, and M. Curado, "Using natural language processing to detect privacy violations in online contracts," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, ser. SAC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 1305–1307. [Online]. Available: <https://doi.org/10.1145/3341105.3375774>
- [36] S. Jugran, A. Kumar, B. S. Tyagi, and V. Anand, "Extractive automatic text summarization using spacy in python & nlp," in *2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 2021, pp. 582–585.
- [37] A. V. Aho, "Algorithms for finding patterns in strings, handbook of theoretical computer science (vol. a): algorithms and complexity," 1991.

# Improving Arabic Cognitive Distortion Classification in Twitter using BERTopic

Fatima Alhaj<sup>1</sup>  
The University of Jordan,  
Amman, Jordan

Ali Al-Haj<sup>2</sup>  
University of Suffolk  
Ipswich, United Kingdom

Ahmad Sharieh<sup>3</sup>, Riad Jabri<sup>4</sup>  
The University of Jordan,  
Amman, Jordan

**Abstract**—Social media platforms allow users to share thoughts, experiences, and beliefs. These platforms represent a rich resource for natural language processing techniques to make inferences in the context of cognitive psychology. Some inaccurate and biased thinking patterns are defined as cognitive distortions. Detecting these distortions helps users restructure how to perceive thoughts in a healthier way. This paper proposed a machine learning-based approach to improve cognitive distortions' classification of the Arabic content over Twitter. One of the challenges that face this task is the text shortness, which results in a sparsity of co-occurrence patterns and a lack of context information (semantic features). The proposed approach enriches text representation by defining the latent topics within tweets. Although classification is a supervised learning concept, the enrichment step uses unsupervised learning. The proposed algorithm utilizes a transformer-based topic modeling (BERTopic). It employs two types of document representations and performs averaging and concatenation to produce contextual topic embeddings. A comparative analysis of F1-score, precision, recall, and accuracy is presented. The experimental results demonstrate that our enriched representation outperformed the baseline models by different rates. These encouraging results suggest that using latent topic distribution, obtained from the BERTopic technique, can improve the classifier's ability to distinguish between different CD categories.

**Keywords**—Arabic tweets; cognitive distortions' classification; machine learning; social media; supervised learning; unsupervised learning; transformers; BERTopic; topic modeling

## I. INTRODUCTION

Researchers attempt to understand the psychological well-being of users by analyzing the published social media content [1], [2]. This content may hold negative, and inaccurate conclusions called cognitive distortions (CDs). Research showed that CDs could be detected in social media [3]. In cognitive psychology, CDs are biased perspectives, patterns, and beliefs that have been reinforced over the years [4]. Usually, individuals diagnosed with depression express higher levels of distorted thinking than others [5], which affect the content they share.

Moreover, CDs can lead to poor behavior and chronic anxiety and are related to symptoms of depression [6]. Examples of these categories are overgeneralization, emotional reasoning, and catastrophizing. Until being diagnosed, a CD can be so problematic to change. Machine learning (ML) and natural language processing (NLP) can improve mental health care by defining CDs within a text. Classifying CDs can be beneficial to the therapy process in two ways. It helps in replacing these inaccurate thoughts and evaluating the improvement over time.

While many previous works with CDs showed promising results with detecting CD, they faced disappointing results with the multi-class CD classification [7], [4], [8]. This work aims to improve the classification task by dealing with the text shortness problem. In general, many classification tasks working with short text fails to achieve high performance according to the sparse representation of the textual data [9], [10], [11]. Extra contextual information can be deployed to overcome the sparse representation and make the data more related, comprehensively expressed, and expand the efficiency of classifiers to handle unseen data.

Contextual topic modeling (TM) can provide extra latent information and identify semantically similar groups in the corpus. The topic-based embeddings can learn global semantics from the entire corpus [12], which provides the features with an opportunity to become more representative.

Also, while analyzing textual cognitive distortions, we noticed that individuals mostly use a common CD when expressing their thoughts about a specific topic or domain. For example, catastrophizing is highly related to relations, self-image terms, and academic achievement. General terms like life, humans, country and government are mostly related to over-generalizing and labeling. This observation encouraged us to utilize TM as extra contextual information in CD classification.

The TM algorithm used in this work is a transformer-based algorithm that utilizes the pre-trained knowledge in the modeling process to provide a topic distribution. So, our primary contribution in this paper is using the state-of-the-art transformer-based TM algorithm to enrich the CD representations and improve its classification task by exploring potential semantically meaningful categories in the data.

The remainder of the paper is structured as follows. In Section II, a summary of related work is presented. Section III overviews some background. Section IV presents the design and methodology used in the research, while the experimental results are analyzed in Section V. Finally, Section VI discuss conclusions.

## II. RELATED WORK

The proposed approach utilizes the TM algorithm to overcome the shortness of social media text to improve CD classifications. Accordingly, this research is related to three main topics; CD classification, short text classification, and TM in text classification.

TABLE I. A SUMMARIZED SAMPLE OF CD-RELATED STUDIES

| Ref. | Strength                                                                                                                                                                                                                                                                                                            | Weakness                                                                                                                                                                                                                                                                                                                                                            |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| [4]  | Using a representative dataset collected from multiple resources; crowdsourcing volunteers, mental health patients, and online therapy programs.                                                                                                                                                                    | Authors reported F-score equal to 0.45, but the trained model cannot predict 7 out of 15 distortions.                                                                                                                                                                                                                                                               |
| [7]  | <ul style="list-style-type: none"><li>• Detecting and classifying CD from patient-therapist interactions.</li><li>• Used a publicly available dataset called: Therapist Q&amp;A, obtained from the Kaggle data science repository.</li><li>• Represent text using different types of linguistic features.</li></ul> | <ul style="list-style-type: none"><li>• Poor inter-annotator agreement.</li><li>• Unclear distinction between different distortions.</li><li>• Misclassification according to the presence of multiple types of distortions in a single text.</li><li>• Detection of CD type fails to yield good results. Weighted F1-score did not score more than 0.30.</li></ul> |
| [13] | The proposed technology determines whether or not the input statement includes cognitive distortions (binary classification). When detecting distortion in a statement, the approach uses "cognitive restructuring" to reframe the statement to repair the distortion.                                              | <ul style="list-style-type: none"><li>• Authors did not utilize ML methods. Participants were asked to classify statements as distorted or undistorted.</li><li>• The study only evaluates the empathy represented in structured and unstructured responses.</li></ul>                                                                                              |
| [14] | <ul style="list-style-type: none"><li>• Uses a large scale of real-world materials from daily narration and diaries from books and web pages in the cognitive-behavioral therapy domain.</li><li>• Utilizing deep learning techniques like word2vec and CNN.</li></ul>                                              | <ul style="list-style-type: none"><li>• The dataset is not available publicly.</li><li>• The approach was described without experimental results or evaluation techniques.</li></ul>                                                                                                                                                                                |
| [8]  | Proved the possibility of detecting cognitive distortions (binary classification) automatically from personal blogs with relatively good accuracy (73.0%).                                                                                                                                                          | Experiments showed a poor performance in multi-class classification, which was justified by the limited size of individual posts and the overall dataset.                                                                                                                                                                                                           |

Researchers applied ML and NLP to look for mental health and cognitive psychology inferences. Many attempts were made to detect and classify a defined set of cognitive distortions in the English language. Table I summarizes a sample of this studies and describes the related advantages and limitations. Most of the proposed work in Table I couldn't classify CD efficiently. Besides, datasets were collected from different resources, and most of them are not publicly available. The proposed approach creates a dataset by crawling social media to overcome this constraint. Also, we utilized the pre-trained bidirectional encoder representations from transformers (BERT) model to improve classification performance.

The limited number of words in tweets leads to sparse co-occurrence patterns, making the classification task more challenging. To tackle the feature sparseness, works in the literature choose between two main approaches; either represent texts in a lower-dimensional space [9] or add external, implicit, and valuable information to enhance the data representation on the feature space [10], [11]. The second method works on enriching the representation of a short text using additional knowledge or semantics [15]. These semantics could be derived internally [16], or from an external resource, such as a collection of longer documents in a similar domain, or a huge resource such as Wikipedia and WordNet [16], [17], [18]. While the mentioned approaches require additional datasets to enrich the text representation, the proposed approach analyzes the topic distribution within the same dataset.

TM methods have been used to improve text classification in the literature. Anantharaman et al. [19] surveyed a set of TM algorithms for classifying short text and large text (document). Results suggested that the latent semantic analysis method (LSA) is the best for short text classification tasks, while latent dirichlet allocation (LDA) performs better on document classification. In addition, Albalawi et al. [20] investigated TM methods by comparing five frequently used methods. Methods were applied to short textual social data to show their capabilities in defining key and meaningful topics. Their study indicates that TM can overcome the noisy data contained in a short text. Also, an efficient application by Phan et al. [21] classified short medical text by exploiting the

hidden topics in large-scale data collections (Wikipedia and MEDLINE datasets).

Furthermore, Dai et al. [22] extended the bag-of-words representation and introduced a new feature space by using a hierarchical clustering algorithm. Yajian et al. [23] worked on extending the short text to a reasonably long one. In their approach, weighted synonyms and related words are generated for every word by the Word2vec and LDA model.

Most proposed approaches use traditional TM algorithms, such as LDA, a word co-occurrence-based method that only works efficiently with long text. However, the previous methods improved the classification results to some extent but still leave considerable space for improvement.

The proposed approach considers the shortness of social media text in CD classification, which (to the best of our knowledge) is not explored in the literature. The novelty of this approach is based on employing a BERT-based model (BERTopic) for the TM, which provides a better contextual representation. The BERTopic model was employed previously for Arabic text modeling in [24]. It showed a better performance than non-negative matrix factorization (NMF) and LDA.

### III. BACKGROUND

#### A. Cognitive Distortions

Social media textual content represents a rich resource for natural language processing techniques to make inferences in cognitive psychology. Machine learning approaches can be applied to learn the distorted thinking patterns to perform CD detection and classification. Table II shows the five considered CDs used in this work.

#### B. Term Frequency-Inverse Document Frequency

One of the weighting methods for feature-based representation is term frequency-inverse document frequency (TF-IDF). It provides a weight for each word. A TF-IDF value is determined by the relative frequency of a word in a specific text and the inverse proportion of the word over the entire corpus, which reflects how relevant a word is to a particular

TABLE II. COGNITIVE DISTORTION AND THEIR DEFINITIONS

| Cognitive distortion | Definition                                                                           |
|----------------------|--------------------------------------------------------------------------------------|
| Inflexibility        | "Having rigid beliefs about how things or people must or ought to be [25].           |
| Over-generalizing    | Creation negative conclusions without full evidence [5].                             |
| Labeling             | To describe self or others negatively without giving credit to counterevidence [25]. |
| Emotional reasoning  | Depend on feelings as a source of facts [5].                                         |
| Catastrophizing      | Enlarging negative statements or events into disasters [25].                         |

text. The TF-IDF is given by the equation 1. Where the weight of the word  $i$  in the text (document)  $j$  is  $w_{ij}$ .  $N$  is the number of documents, and  $tf_{ij}$  is the frequency of the word  $i$  in the document  $j$ , and  $df_i$  is the number of documents that contain the word  $i$  [26].

$$W_{ij} = tf_{ij} * \log \frac{N}{df_i} \quad (1)$$

### C. Topic Modeling with BERT

Topic modeling (TM) is an analytical unsupervised learning model that discovers topics distribution in a corpus. In this context, a topic can be defined as a repeated pattern of terms [27]. Mainly, the TM algorithm provides two distributions; document-topic and topic-term.

Recently, a transformer-based algorithm was deployed for TM called BERTopic [28]. The algorithm uses three primary phases to produce a topic's distribution for a set of documents. First, it creates sentence embedding. Second, it creates clusters of semantically similar sentences. The last step includes creating topic representation with c-TF-IDF. Each step will be elaborated on in the methodology section.

The usual TF-IDF equation defines the importance of a word between different documents. Differently, the class-based TF-IDF (c-TF-IDF) [28] defines the importance of a word within a class (topic). It treats all documents in a single class as a single document. Equation (2) finds the c-TF-IDF of each word, where  $W_{ic}$  is the weight of word  $i$  in class  $c$ . The frequency of word  $i$  in class  $c$  is  $tf_{ic}$ .  $A$  is the average number of a word per class, and  $f_i$  is the frequency of word  $i$  across all classes.

$$W_{ic} = tf_{ic} \times \log\left(1 + \frac{A}{f_i}\right) \quad (2)$$

Essentially, the BERTopic technique utilized two Algorithms; UMAP and HDBSCAN. Uniform manifold approximation and projection (UMAP) [29] is an algorithm for non-linear dimension reduction that combines manifold learning and topological data analysis. The algorithm optimizes a low-dimensional graph to be structurally similar to the original graph. While HDBSCAN algorithm [30] is a hierarchical clustering algorithm that stands for hierarchical density-based spatial clustering of applications with noise. The algorithm first transforms the space according to the density or sparsity of the data to provide a distance-weighted graph, and then it creates the minimum spanning tree for this graph. The algorithm also represents the connected components by a cluster hierarchy

then condenses them based on the minimum cluster size parameter. Finally, it extracts clusters from the condensed tree.

## IV. METHODOLOGY

In general, usual text mining methods have some limitations in classifying short texts. The limitations are related to the sparsity of co-occurrence patterns in short texts, and the lack of context information (semantic features) [31]. A common method to overcome these limitations is to enrich the texts with additional information to make data more related and extend the classifiers' coverage to perform better in future data [21]. The additional information can be directly attached to the textual representation. After that, the standard classification is performed.

The proposed approach in this paper contains two phases; the enrichment phase and the classification phase. Two main steps are performed in the enrichment phase; the first step is acquiring the additional information, while the second step combines them with the original data representation.

In this work, the additional information source is the output of a TM algorithm (described briefly in the background section). The algorithm generates probability distributions of the hidden topics in the corpus. BERTopic [28] is used as a TM technique, where it takes advantage of BERT embeddings. The BERTopic technique produces the distribution of topics within a document  $d$ . Later, this distribution (BERTopic ( $doc_d$ )) is combined with text embedding.

Fig. 1 shows an overview of the proposed approach that uses BERTopic processes to enrich the document representation with the hidden topic distribution. The following steps summarize the BERTopic processes:

- 1) Generate embeddings for tweets to provide numerical vector representations, where each tweet is considered a single document. This step is based on a pre-trained deep bidirectional transformer. This type of embedding is very powerful for language understanding and can capture the semantic relations between words.
- 2) Reduce the dimensionality of the embedding vectors to create a lower-dimensional embedding of tweets by using the UMAP algorithm [29]. The algorithm preserves the local and global structure in lower dimensionality. Cosine-similarity is used as a distance metric to measure distances between data points.
- 3) Perform clustering using HDBSCAN [30] to find highly similar text in the semantic space that produces a single topic. Defining the number of topics in this stage is a crucial decision. It is supposed to be reasonable and compatible with the dataset size and topics-related words. The generated topics can be too fine-grained, which results in a massive number of topics. Also, it can dismiss important details by defining a small number of clusters. In BERTopic, the algorithm infers the ideal number from the data itself. This step applies iterative merging for the most similar topics. Cosine

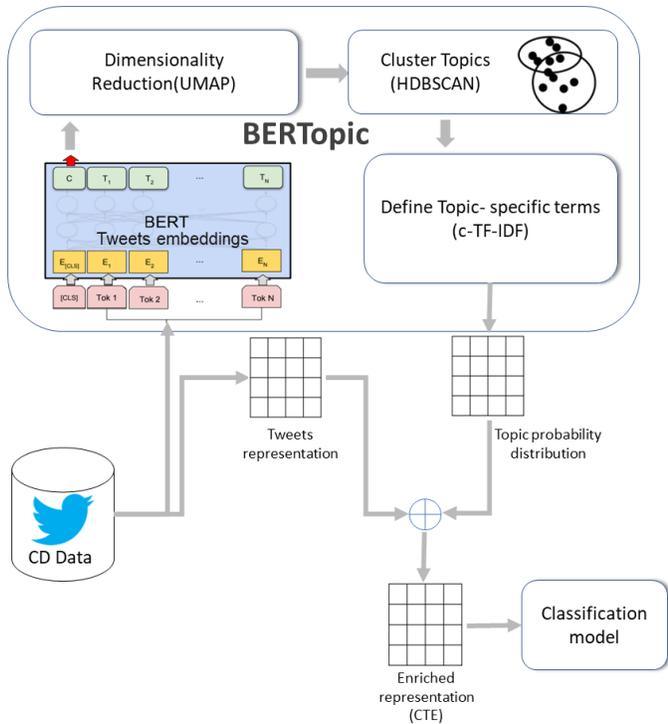


Fig. 1. A Graphical Representation of the TM Enrichment Approach for CD Classification.

similarity between c-TF-IDF vectors is utilized in this process.

- 4) Define the essential words in a topic according to the c-TF-IDF equation. Words with a high score are more representative of the topic. Accordingly, every tweet is related to a set of topics with different probabilities. In this step, BERTopic generates a topic distribution vector for each tweet,  $Pr(topic | tweet)$ . A graphical representation map for this process is shown in Fig. 2.

The previous processes produce topic distribution for each document as  $BERTopic(doc_d)$ . Then, the distribution is used in Algorithm 1 to produce the final contextual topic embedding (CTE) for each document. Mainly, the algorithm uses two types of document representations; topic distribution and word embedding. Then, it performs averaging and concatenation. The nested for (line 2-4) creates word embeddings for each word in a document  $d$  using the word2vec model, while line 5 constructs document embeddings. All word vectors in document  $d$  are averaged to produce a single vector in the same embedding space.

Combining BERTopic distribution and word2vec representation keeps the semantic information and creates contextual topic representation. The topic representation and the word embedding are composed in line 6 to produce CTE for a given document ( $d$ ), where  $\oplus$  represents the arithmetic concatenation operator. We use "featureUnion" function<sup>1</sup> in Python to

<sup>1</sup><https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.FeatureUnion.html>

perform the concatenation.

---

**Algorithm 1:** Generate Contextual Topic Embedding

---

**input :**  $M$ : Maximum sequence of words,  
 $D$ : Number of documents,  
 $doc_d = \{w_1, \dots, w_M\}$ .  
 $BERTopic(doc_d)$

**output:**  $CTE(doc_d)$

- 1 **for**  $d \in \{1, \dots, D\}$  **do**
- 2     **for**  $n \in \{1, \dots, M\}$  **do**
- 3         | Embedding( $w_{dn}$ )= word2vec( $w_{dn}$ )
- 4     **end**
- 5     Embedding( $doc_d$ )= Avg(Embedding( $w_{d1}$ ),...,Embedding( $w_{dm}$ ))
- 6      $CTE(doc_d)$ =Embedding( $doc_d$ ) $\oplus$   $BERTopic(doc_d)$
- 7 **end**

---

The next step is feeding the enriched representation  $CTE(doc_d)$  to different classifiers and evaluating the results. Because this approach uses hidden topic knowledge of the corpus as the primary source of enrichment, its effectiveness is corpus-dependent.

## V. EXPERIMENTAL RESULTS AND DISCUSSION

The dataset was collected by crawling Twitter data. First, a translated version of the cognitive distortion scale [32] was utilized. Volunteers were asked to express distorted thoughts they had. The initial responses were used to define a keyword list for each category. Keywords were defined as the top frequent and influential words in the responses. For example, one of the CD patterns is inflexibility. The related seed words were: (لازم / يجب / لابد / واجب / مفروض) / (مفترض / محتم) , which mean: (obligatory / must / must / obligatory / obligatory / unavoidable / supposed) respectively. The defined seed words were used for crawling Twitter by streaming API in June 2021.

Tweets were labeled by two reliable annotators who work in the psychotherapy domain. The annotators labeled the data independently of one another. Inter annotator agreement (IAA) was calculated to evaluate the annotation quality. Cohen's kappa coefficient between annotators was 0.817, which indicates an almost perfect agreement according to Kappa's interpretations [33]. We only considered texts that both annotators labeled with the same label (9250).

We excluded the noise usually found in a social media text. The preprocessing operation included removing Arabic diacritics, Arabic and English punctuations, stop words, emojis, and non-Arabic characters. Also, The Arabic letters were normalized, where we replaced letters with different shapes with one of its defined shapes. "Alef" (آ , ا , إ) into ا, and "Ta Marbouta" (ة) into (o). In addition, we removed tweets-related noise like users' mentions, usernames, hashtags, and Twitter handles (@user), RT, <>. The last stage in the preprocessing is stemming, which removes any suffixes, prefixes, or infixes from words. This step reduces the derived or inflected words

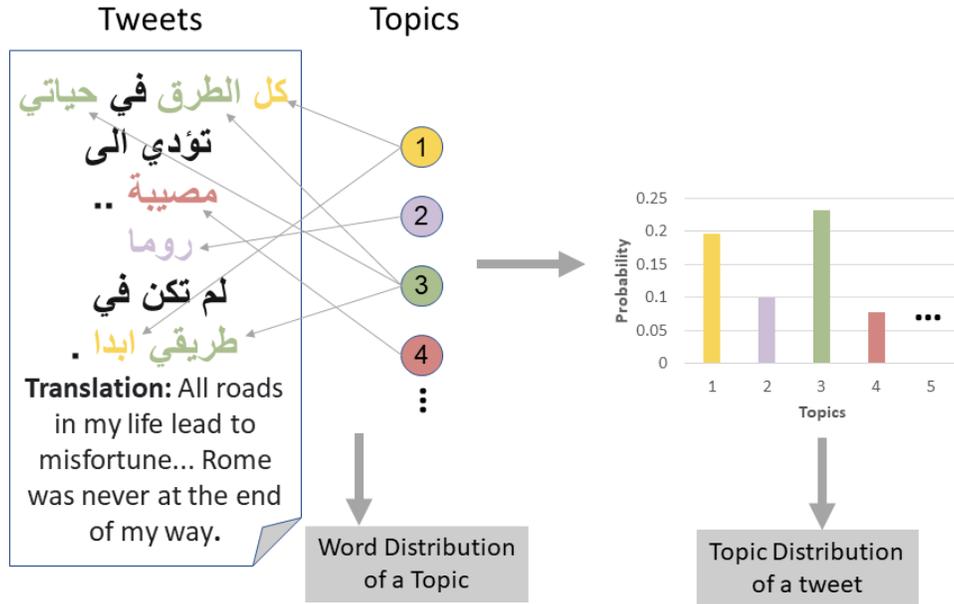


Fig. 2. A Graphical Representation for Generating Topic Distribution Vectors.



Fig. 3. Bar Charts for Top c-TF-IDF Scores in Six Topics.

TABLE III. A PERFORMANCE COMPARISON OF MULTIPLE CLASSIFIERS, USING BOTH WORD2VEC AND THE ENRICHED FEATURES (CTE)

| Model    | Word2vec |          |           |        | Enriched Representation (CTE) |          |           |        |
|----------|----------|----------|-----------|--------|-------------------------------|----------|-----------|--------|
|          | F1-score | Accuracy | Precision | Recall | F1-score                      | Accuracy | Precision | Recall |
| DT       | 0.6329   | 0.6329   | 1         | 0.6333 | 0.9667                        | 0.9667   | 0.9667    | 0.9667 |
| SVM      | 0.9273   | 0.9274   | 1         | 0.9268 | 0.9792                        | 0.9792   | 0.9792    | 0.9792 |
| RF       | 0.8372   | 0.8632   | 1         | 0.8644 | 0.9004                        | 0.9004   | 0.9004    | 0.9004 |
| KNN      | 0.7701   | 0.7714   | 1         | 0.7911 | 0.8640                        | 0.8641   | 1         | 0.8641 |
| XGBoost  | 0.8789   | 0.8782   | 1         | 0.8797 | 0.9823                        | 0.9823   | 0.9823    | 0.9823 |
| Bagging  | 0.7760   | 0.7798   | 1         | 0.7771 | 0.9710                        | 0.9709   | 0.9709    | 0.9709 |
| Stacking | 0.9280   | 0.9297   | 1         | 0.9270 | 0.9710                        | 0.9813   | 0.9710    | 0.9710 |

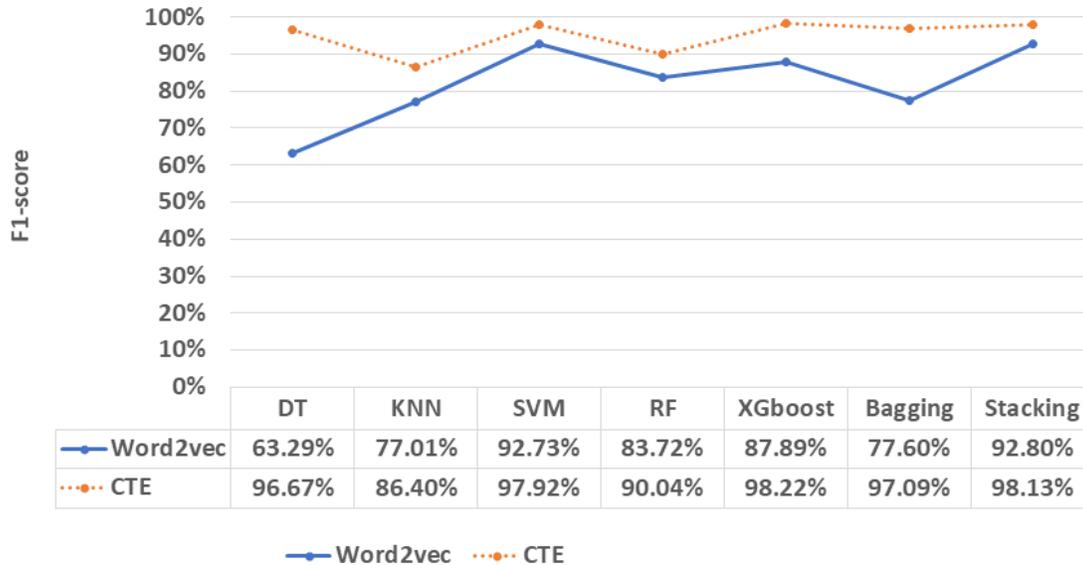


Fig. 4. F1-score Comparison between Classifiers that used Word2vec only and Classifiers that Utilized the Enriched Features (CTE).

into their related stems and relates all the variations into one stem. We used Tashaphyne<sup>2</sup>, an Arabic light stemmer tool.

All experiments were applied by splitting data randomly into 75% (6,940 tweets) for training and 25% (2,310 tweets) for testing. Ten runs were performed for each experiment, and we reported the average results to overcome the performance variation problem. We used the NVIDIA Tesla K80 GPU provided on the Google colaboratory notebook.

The word2vec representation and BERTopic [28] algorithm are employed in this approach. We applied the BERTopic algorithm on the preprocessed CD dataset using Python’s ”bertopic” library. Arabert model [34] was used as a pre-trained language model. Each term in the corpus contributes differently to each topic. Fig. 3 shows the bar charts of the first six topics with a subset of the most contributing terms according to the c-TF-IDF scores.

We reduced the number of topics by starting from the least frequent topic and merging topics as long as the similarity exceeds 0.915, the threshold suggested in the BERTopic documentation. The number of topics was reduced from 70 topics to 47. Afterward, the resulting topic distributions for tweets in the dataset were concatenated to the word2vec representation.

The resulting enriched features are fed into six classifiers.

Table III demonstrates the results of four performance metrics; macro-accuracy, macro-precision, macro-recall, and macro-F1. Also, Fig. 4 shows F1-scores of classification using enriched features approach compared with classification using word2vec features. According to Fig. 4, the performance of the proposed approach improved all the classification results. DT and bagging results boosted the most. These encouraging results suggest that using latent topic distribution, obtained from BERTopic, can improve the classifier’s ability to distinguish between different CD categories. The baseline classifiers could not fully capture the relations of a CD category and a subset of topics due to the shortness of tweets. On the contrary, the enriched features emphasized these relations.

## VI. CONCLUSION

The constant evolution in natural language processing and machine learning has made it possible to address different cognitive distortions in a text. This paper suggests an approach for enriching short textual representations to improve cognitive distortions’ classification of the Arabic context over Twitter. The enrichment approach considers the shortness of social media text and the sparsity of word co-occurrence patterns, which (to the best of our knowledge) is not explored in the literature. It also utilizes a transformer-based topic modeling algorithm (BERTopic) that employs a pre-trained language model (AraBERT). For evaluation, various

<sup>2</sup><https://pypi.org/project/Tashaphyne/>

classifiers were used; DT, k-NN, SVM, RF, XGBoost, stacking and bagging. Experimental results showed that the proposed approach improved the performance of the defined classifiers with different rates using an Arabic CD dataset. We will study other TM approaches with advanced feature representations and embeddings for future work. In addition, CD data from other sources and languages can be tested to examine the generality of the proposed approach. As a recommendation for other researchers, we encourage examining transformer-based topic modeling to empower different supervised tasks in NLP. Also, it is good practice to define an approximation for topics' granularity in different domains.

## REFERENCES

- [1] D. D'Hotman and E. Loh, "Ai enabled suicide prediction tools: a qualitative narrative review," *BMJ Health & Care Informatics*, vol. 27, no. 3, 2020.
- [2] A. Wongkoblap, M. A. Vadillo, and V. Curcin, "Social media big data analysis for mental health research," in *Mental Health in a Digital World*. Elsevier, 2022, pp. 109–143.
- [3] D. J. Dozois, "Depressive cognition on twitter," *Nature Human Behaviour*, vol. 5, no. 4, pp. 414–415, 2021.
- [4] B. Shickel, S. Siegel, M. Heesacker, S. Benton, and P. Rashidi, "Automatic detection and classification of cognitive distortions in mental health text," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, 2020, pp. 275–280.
- [5] K. C. Bathina, M. Ten Thij, L. Lorenzo-Luaces, L. A. Rutter, and J. Bollen, "Individuals with depression express more distorted thinking on social media," *Nature Human Behaviour*, vol. 5, no. 4, pp. 458–466, 2021.
- [6] A. Barak and J. M. Grohol, "Current and future trends in internet-supported mental health interventions," *Journal of Technology in Human Services*, vol. 29, no. 3, pp. 155–196, 2011.
- [7] S. Shreevastava and P. Foltz, "Detecting cognitive distortions from patient-therapist interactions," in *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 2021, pp. 151–158.
- [8] T. Simms, C. Ramstedt, M. Rich, M. Richards, T. Martinez, and C. Giraud-Carrier, "Detecting cognitive distortions through machine learning text analytics," in *2017 IEEE international conference on healthcare informatics (ICHI)*. IEEE, 2017, pp. 508–512.
- [9] H. M. K. Kumar and B. S. Harish, "A new feature selection method for sentiment analysis in short text," *Journal of Intelligent Systems*, vol. 29, no. 1, pp. 1122–1134, 2020. [Online]. Available: <https://doi.org/10.1515/jisys-2018-0171>
- [10] A. Bagheri, A. Sammani, P. G. van der Heijden, F. W. Asselbergs, and D. L. Oberski, "Etm: Enrichment by topic modeling for automated clinical sentence classification to detect patients' disease history," *Journal of Intelligent Information Systems*, vol. 55, no. 2, pp. 329–349, 2020.
- [11] D. Bollegala, V. Atanasov, T. Maehara, and K.-i. Kawarabayashi, "Classinet—predicting missing features for short-text classification," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 12, no. 5, pp. 1–29, 2018.
- [12] S. Seifollahi, M. Piccardi, and A. Jolfaei, "An embedding-based topic model for document classification," *Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 3, pp. 1–13, 2021.
- [13] R. R. Morris and R. Picard, "Crowdsourcing collective emotional intelligence," *arXiv preprint arXiv:1204.3481*, 2012.
- [14] X. Zhao<sup>12</sup>, C. Miao<sup>12</sup>, and Z. Xing, "Identifying cognitive distortion by convolutional neural network based text classification," *International Journal of Information Technology*, vol. 23, no. 1, 2017.
- [15] A. Sun, "Short text classification using very few words," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012, pp. 1145–1146.
- [16] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, "Exploiting internal and external semantics for the clustering of short texts using world knowledge," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, pp. 919–928.
- [17] S. Zelikovitz, W. W. Cohen, and H. Hirsh, "Extending whirl with background knowledge for improved text classification," *Information Retrieval*, vol. 10, no. 1, pp. 35–67, 2007.
- [18] M. Chen, X. Jin, and D. Shen, "Short text classification improved by learning multi-granularity topics," in *Twenty-second international joint conference on artificial intelligence*, 2011.
- [19] A. Anantharaman, A. Jadya, C. T. S. Siri, B. N. Adikar, and B. Mohan, "Performance evaluation of topic modeling algorithms for text classification," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2019, pp. 704–708.
- [20] R. Albalawi, T. H. Yeap, and M. Benyoucef, "Using topic modeling methods for short-text data: A comparative analysis," *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020.
- [21] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 91–100.
- [22] Z. Dai, A. Sun, and X.-Y. Liu, "Crest: Cluster-based representation enrichment for short text classification," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 256–267.
- [23] Z. Yajian, D. Dingpeng, and C. Junhui, "A short text classification algorithm based on semantic extension," *Chinese Journal of Electronics*, vol. 30, no. 1, pp. 153–159, 2021.
- [24] A. Abuzayed and H. Al-Khalifa, "Bert for arabic topic modeling: An experimental study on bertopic technique," *Procedia Computer Science*, vol. 189, pp. 191–194, 2021, aI in Computational Linguistics. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1877050921012199>
- [25] L. Rojas-Barahona, B.-H. Tseng, Y. Dai, C. Mansfield, O. Ramadan, S. Ultes, M. Crawford, and M. Gasic, "Deep learning for language understanding of mental health concepts derived from cognitive behavioural therapy," *arXiv preprint arXiv:1809.00640*, 2018.
- [26] W. Zhang, T. Yoshida, and X. Tang, "A comparative study of tf\*idf, lsi and multi-words for text classification," *Expert Systems with Applications*, vol. 38, no. 3, pp. 2758–2765, 2011.
- [27] V. K. Gunjan, J. M. Zurada, B. Raman, and G. Gangadharan, *Modern Approaches in Machine Learning and Cognitive Science: A Walk-through*. Springer, 2020.
- [28] M. Grootendorst, "Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics." 2020. [Online]. Available: <https://doi.org/10.5281/zenodo.4381785>
- [29] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.
- [30] R. J. Campello, D. Moulavi, and J. Sander, "Density-based clustering based on hierarchical density estimates," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2013, pp. 160–172.
- [31] L. Yang, C. Li, Q. Ding, and L. Li, "Combining lexical and semantic features for short text classification," *Procedia Computer Science*, vol. 22, pp. 78–86, 2013.
- [32] R. Covin, D. J. Dozois, A. Ogniewicz, and P. M. Seeds, "Measuring cognitive errors: Initial development of the cognitive distortions scale (cds)," *International Journal of Cognitive Therapy*, vol. 4, no. 3, pp. 297–322, 2011.
- [33] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
- [34] W. Antoun, F. Baly, and H. Hajj, "Arabert: Transformer-based model for arabic language understanding," *arXiv preprint arXiv:2003.00104*, 2020.

# Investigation Framework for Cloud Forensics using Dynamic Genetic-based Clustering

Mohammed Y. Alkhanafseh<sup>1</sup>  
Department of Computer Science  
Birzeit University  
Ramallah, Palestine

Mohammad Qataweh<sup>2</sup>  
Department of Computer Science, KASIT  
The University of Jordan  
Amman, Jordan

Wesam Almobaideen<sup>3</sup>  
EE and Computing Sciences Departmen  
Rochester Institute of Technology-Dubai  
Dubai, UAE

**Abstract**—Cloud computing allows a pool of resources, such as storage, computation power, communication bandwidth, to be shared and accessed by many users from different locations. High dependency on sharing resources among different cloud users allows some attacker to hide and commit crimes using cloud resources and as a result, cloud computing forensics become essential. Many solutions and frameworks for cloud computing forensics have been developed to deal with cloud based crimes. However, many problems and issues face the proposed solutions and frameworks. In this paper, a new framework for cloud computing forensics is proposed to enhance the investigation process performance and accuracy by adding a new stage to conventional stages. This new stage includes the implementation of a new way for matching based on the LSH algorithm. The proposed framework evaluation results show an improvement for matching and accurate cluster retrieval through the collection process.

**Keywords**—Cloud computing forensics; genetic clustering algorithms; genetic dynamic clustering; forensics framework; digital forensics

## I. INTRODUCTION

Digital forensics is a science which is defined by many researchers, see [31][20], as a science of retrieving, examining, and analysing digital evidences collected from digital devices. A Digital evidence, according to [20], is defined as the information and data of investigative value that is stored on, received, or transmitted by a digital device. The goal of digital forensics is to detect and extract the evidences, analyze the collected evidences and preserve related ones in a format that can be presented in a court [24][18].

Cloud computing is one of the most popular technologies since different users and companies around the world are rapidly becoming more dependent on cloud computing. This means that millions or even billions of files have been uploaded to cloud computing resources. Security of resources over the cloud is a challenging issue that has to be addressed [43] [3]. Fog computing is similar to cloud computing but closer to the user which also has many security concerns [2]

Cloud computing forensics refers to one of the digital forensics branches This branch has gained popularity from the importance of today's cloud computing. Many researchers describe cloud computing forensics as an application for digital forensics for cloud computing resources to investigate the crimes that occurred in cloud computing resources. Text files and textual information represent a very high percentage of types of evidence, particularly in cloud computing [20].

Accordingly, and due to the extra high volume of data stored on the cloud, an enhanced solution is needed to handle textual evidence by extracting and encouraging the handling of textual evidence.

In recent years, the number of digital crimes that involve internet and computer has grown, which has encouraged a lot of companies to include measures in their products to assist law enforcement in using digital evidences to determine the perpetrators, methods used, timing, and victims of computer crimes. The process of applying digital forensics in the huge volume of files which stored in cloud computing resources is very hard. Therefore, the idea of clustering has been applied to facilitate the investigation process. Clustering is the process of grouping objects or documents related to each other in the same group. This is important in the investigation process, since the investigator needs to search for all documents in the storage space to define if there are any documents related to the original file [17].

This paper presents a new approach to cloud computing forensics, which solves some of existing problems and challenges. The proposed solution consists of two main parts. The contribution of this paper can be summarized in the following points:

- A pre-investigation stage has been added to the standard forensics stages as part of the propose Cluster Based Investigation Framework (CBIF). Through this stage, a new hierarchical clustering algorithm capable of handling all types of evidence is added.
- Genetic Based Dynamic Clustering Algorithm (GB-DCA) has been developed to divide a dataset into suitable number of clusters in cloud storage.
- A new search and matching technique based on using a locality-sensitive hashing algorithm has (LSH) to enhance the accuracy of the matching process.
- Additionally, the proposed idea can match parts of the files based on the concept of hierarchical matching and examining process.

This paper is organized as follows. Section 2 brings forth the related work. Section 3 presents a brief description of cloud computing and cloud forensics. Section 4 illustrates the proposed idea and foremost step of the proposed solution. Section 5 details the experimental implementation and results from the discussion. We conclude with Section 6.

## II. RELATED WORK

In the field of digital forensics, some frameworks have been suggested as a general frameworks while others were proposed for lower levels of digital forensics branches. Example of frameworks that are suggested as generic framework for digital forensics is an integrated digital forensics process model [41], which is a standard framework for digital forensics and can apply to all branches of digital forensics. This framework contains all standard stages of digital forensics with some changes in the level of combining evidence collection and the preservation stage into a single stage to make sure the identified evidence does not change through the collection stage.

Another model proposed as a generic framework is the integrated framework for digital forensics [48], where the proposed model is a standard model that helps the investigators to follow a uniform approach in digital forensics. The model contains all conventional stages of digital forensics. This framework is classified as a simple framework based on [31]. And other models that fall under general framework such as the model suggested through [40], [8], [32], and [7].

On the level of the computer forensics branch, many frameworks were suggested, such as the computer forensics investigation approach to evidence in cyber-crime [13], which aims to define a new approach to solve and enhance the stage of computer forensics examination. The model meets Italian legislation and could probably be used in other countries.

Numerous frameworks and solutions have been proposed in other digital forensics branches [31], [47]. One instance includes the frameworks and solutions proposed in the IoT forensics branch, which refer to one of the digital forensics branches, such as the one in [13]. The idea of the proposed system relates to using Block-Chain through the investigation process to improve the collection stage's security level and credibility.

The model proposed in [21] refers to a new model for reviewing and investigating cyber attacks, where digital

other framework was proposed in [37], which refers to a new framework for mobile forensics, since a lot of crimes was applied on the resources of mobile devices, specially with rapid development of wireless network and smart mobile. The idea that proposed here refers to depends on the clustering process to facilitate the investigation as all.

The model proposed in [6] refers to the first framework, which contains various safety principles proposed by the ISO standard. Another framework for investigating crimes committed in IoT devices is the IoT forensics framework for the smart environment [39]. The proposed model was specified to investigate the crimes committed in a smart environment, where the proposed model is a lightweight model that is well equipped to be compatible with IoT resources. Moreover, this model can classify as a generic model for all IoT crimes with a high level of privacy based on the principles mentioned for this point.

Other frameworks and solutions proposed in the other branches of digital forensics include the framework and solutions proposed to investigate the crimes committed in a cloud

computing environment. One of these frameworks is presented in [26], where the main contribution for the proposed model is defining the difference between the evidence collection stage and the preservation stage, since many frameworks are merged between the two stages. In [42], a new framework for cloud computing with a fundamental change in the stages of conventional frameworks is illustrated. The change affects the proposed model of the identification and collection stages, and the rest of the stages did not change. The science of the proposed framework focuses on the first two stages. Additionally, an open cloud forensics model is also an example of a framework proposed for cloud computing forensics [51]. The framework consists of primary stages of digital forensics with an update on the levels of preservation and collection. The model merges between these two stages in a single stage to make sure that the evidence does not change through the collection stage, and merges between the analysis stage and examination stage in a new stage called the organization stage.

One solution proposed for the cloud forensics process is to secure logging as a service for cloud forensics [50]. The idea of this solution is to introduce secure logging as a service that allows the cloud service provider to store virtual machine logs and to provide access for the investigators while preserving the privacy for the tenant of the cloud. In [35], the goal of the proposed model is to provide as much information about each record, such as when any trigger occurs, as when a new record is added or deleted.

When the cloud computing environment contains multiple tenants for its resources, the privacy and security of evidence are essential. One of the proposed solutions to deal with this problem is presented in [5], where the idea of the proposed solution is to use a third party to check and evaluate data collected by investigators. Another solution to solve the multi-tenant problem is proposed in [4], where the idea depends on upholding the confidentiality and integrity for evidence that used through the investigation process.

Yet an additional issue facing the investigation process in cloud computing is the data gathering challenge, which faces the acquisition stage of the investigation process, because the data is distributed amongst different servers which ultimately decrease the performance of the investigation process. Specific frameworks and solutions have been proposed to deal with this problem, such as the framework proposed by Adams [44], which suggested a new cloud forensics model which consists of a planning stage, on-site survey stage, and acquisition stage. This applies to deal with the process of acquisition without any further intervention.

Numerous frameworks and approaches have been suggested to investigate crimes committed in computer networks. Frameworks that scales for a large-scale environment are called dynamic network forensics and have a specific property for investigation such as the framework presented in [25], which investigates the crimes that occur in network infrastructure. The framework contains the standard stages of any digital forensic framework with additional stages to enhance the investigation process, such as the evidence reduction stage, which can enhance the accuracy of the investigation process. Graphic network forensics have many frameworks are specialized for graphics-based network forensics crimes [49], [28], [38]. Many approaches and solutions have been proposed to

investigate soft-computing crimes specific to network analysis and monitoring. The systems proposed in this field are tailored to the environment that includes many attacks, because this type of network forensics can analyze the data collected and identify relevant attacks [9], [36].

Numerous researchers identified cloud computing forensics [14], as an application for digital forensics in cloud computing as a subset of network forensics. NIST framework [29], defines it as an application of science to the identification, examination, and analysis of data while preserving the integrity of data and maintaining the chain of custody for the data through the investigation process. Any information or data stored or extracted from digital devices can be evidence or part of the evidence; these pieces of evidence are analyzed through the investigation process.

The goal of digital and cloud investigation is to ensure that the collected evidence is admissible in a court of law and to maintain that the chain of custody is essential and documented, and is thus stored throughout the entire investigation process [16]. The chain of custody is the process of recording and storing digital evidence as well as managing historical history to protect the custody chain from any alteration or modification or damage by any unauthorized user, and so, a high level of care is required.

The investigation of crimes occurring in the cloud computing environment is faced with many issues. These issue can be summarized through the Table I as follow

Based on the issues mentioned above, a solution for cloud computing forensics is required, as many of the proposed frameworks do not provide a general solution to the aforementioned problems. The solution suggested in this article aims to avoid all the above mentioned significant issues. The proposed solution can be applied in all branches of digital forensics in addition to cloud computing, such as applying the proposed solution in computer forensics, cloud forensics, network forensics, database forensics, and all others.

TABLE I. KEY ISSUE THAT FACES CLOUD COMPUTING FORENSICS PROCESS.

| Key Issue Description                   | The Reason of Key Issue                                                                                                                                                                                                                                                                        |
|-----------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Searching for Candidate Files           | Cloud computing contains a considerable amount of massively distributed information from different users, which makes the searching for evidence a laborious process                                                                                                                           |
| Privacy of users documents              | Cloud computing contains a considerable amount of massively distributed information from different users, which makes the searching for evidence a laborious process.                                                                                                                          |
| Time Needs for locating the target file | Searching for the target file or any of its pieces is time-consuming, based on the massive amount of files in the data center. On that basis, a new matching technique is required to improve the efficiency and accuracy of the matching process for evidence.                                |
| The Integrity of users data             | In conventional ways, the investigator must check all files to check whether or not they relate to the target file, which may violate all data in the storage center if the investigator is a criminal. A technique is required to check only those files that are related to the target file. |

### III. PROPOSED IDEA

The idea of the proposed solution focuses on resolving shortcomings and problems that face previous related solutions such as these related to the used conventional clustering algorithm in [17] [33]. The solution which proposed in [17] depends on using conventional versions of k-means and k-medoid clustering algorithms during the pre-investigation stage to enhance the accuracy of investigation process. While in [33] the approach suggested is based on the usage of the hierarchical clustering algorithm in the pre-investigation process to enhance the investigation’s accuracy and performance.

Another weak-point faces the previous suggested frameworks is related to the privacy of users’ data through the investigation process. This is because many of proposed framework does not proposed any solution for the privacy problem [1] and [34].

The proposed framework consists of all stages of a conventional framework for digital investigation, with additional stages to achieve enhancement goals. The pre-investigation stage is essential and was added in the proposed framework to enhance the investigation process from in regards with the performance, accuracy, and security. The proposed solution consists of a set of stages as follows:

- Pre-investigation stage: this stage reduces the amount of information submitted to the investigator. The stage can improve the accuracy and efficiency of the investigation process.
- Evidence collection stage: in this stage the investigator is responsible for collecting and transferring to the investigator side only those clusters identified in the preceding stage which are related to the target files.
- Matching stage: conventional matching, such as sequential and carving based matching, is usually used in related framework found in the literature [1]. LSH based matching is suggested in the proposed framework in order to improve the matching process accuracy and efficiency.
- Analysis and presentation stage: in this stage the investigator is responsible for the analysis of the evidence gathered and finalizing the report to be presented in front of the court. .5

The improvement introduced in CBIF framework relates to the addition of the pre-investigation stage and to the improvement of the efficiency and accuracy of the investigation process

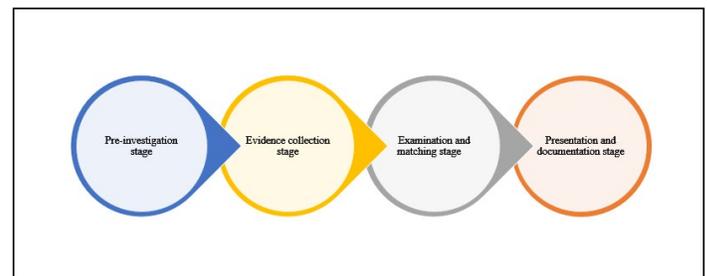


Fig. 1. Main Stages for the Cluster based Investigation Framework(CBIF).

by introducing the LSH based matching. Detailed discussion of each of the CBIF enhanced points will be presented in the subsections below.

A. Hierarchical Clustering used in the Pre-Investigation Stage

The objective of the hierarchical clustering algorithm used in the suggested idea is to improve the investigation process by enhancing the crime-related matching and related data collection within proper clusters. The idea of hierarchical clustering algorithm merges between a static clustering algorithm based on a genetic algorithm [23][27][30], and a dynamic clustering algorithm using an evolutionary algorithm [44][38].

Fig. 1 illustrates how the proposed hierarchical clustering algorithm operates. The algorithm starts by applying a genetic-based static clustering algorithm (GBSCA) to divide the storage center into a set of clusters based on files data-type. The storage center in Fig. 2 is divided into four main clusters; one for audio files, one for video files, one for text files and documents, and one for executable files. Afterward, another clustering round process is applied to each of the clusters formed. The second-level clustering algorithm is a genetic-based dynamic clustering algorithm (GBDCA). The objective of this second round is to divide each cluster generated into a suitable number of sub-clusters. The execution of all steps in Fig. 2 is the responsibility of the cloud service providers.

The evaluation metric which is used in the evaluation process refers to the Sum Square Error(SSE), whereas the equation for this evaluation metric is as follows.

$$SSE = \sum_{i=1}^k \sum_{j=1}^{|c_i|} D(c_j - f_i)^2 \dots (1)$$

In Equation (1) D refers to the Euclidean distance between Centroid file Cj and a specific file fi. The value of SSE must be minimized as much as possible.

B. Genetic based Dynamic Clustering Algorithm (GBDCA)

Dynamic clustering algorithm refers to a clustering strategy used to divide the storage center into a set of clusters [24], which is different from the conventional clustering algorithm. The goal of a dynamic clustering algorithm in CBIF is to divide the storage center into a proper number of clusters. Dynamic clustering algorithm consists of a set of iterations; the

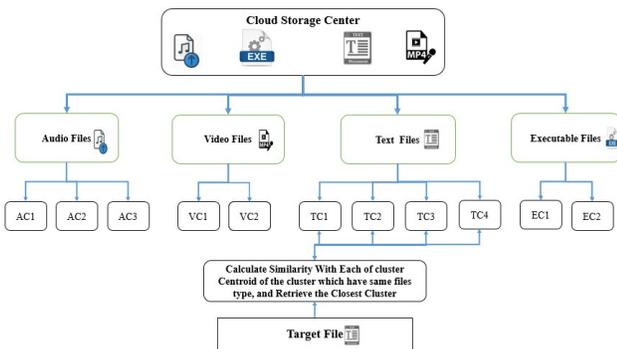


Fig. 2. The Main Steps of the Pre-investigation Step of the Proposed Idea to Divide the Cloud Storage Center into a Suitable Number of Clusters.

maximum number of iterations is defined before the algorithm is started. The iterations begin with an arbitrarily small number of clusters; at each iteration the number of clusters increases by one until the maximum number of iterations is reached. The number of clusters used is evaluated, based on two fitness functions, the internal fitness evaluation, and external fitness evaluation that is measured in each iteration.

The goal of the internal loop is to select the best centroid file which achieves the highest fitness value based on the SSE equation. fitness is achieved based on specific centroid files being compared with the best fitness value.

The results of applying the genetic-based dynamic clustering, as the second level in the hierarchical clustering algorithm, is getting the most suitable set of clusters based on the reported best fitness function of the external loop. The flowchart below shows the main stages of the GBDC algorithm.

The flowchart in Fig. 3 represents the main stages of the dynamic clustering algorithm, which can facilitate the investigation process by dividing the files into a suitable number of clusters. This means that the cluster which achieves a high level of similarity to the target file will contain all the files actually related to the target file indicated by low intra-distance which means high similarity between files in the same cluster. The equation for the ratio between inter-distance, i.e. the differences between files in different clusters, to intra-distance, which is proposed in [12] is mentioned in equation (2). This equation represents the external fitness function proposed in GBDCA, which is responsible for selecting the optimal number of clusters for the files in the original data set. The output of the external fitness evaluation helps in descending on the best number of related clusters:

$$R = \frac{InterDistance}{K - 1} * \frac{IntraDistance}{n - K} \quad (1)$$

...(2)

In equation (2), “R” refers to the ratio between the inter-distance and intra-distance, “K” refers to the maximum number of clusters that are allowed in a solution, and “n” refers to the number of data instances, e.g. number of files. In the Algorithm 1, the similarity between data centers and all files in the cloud

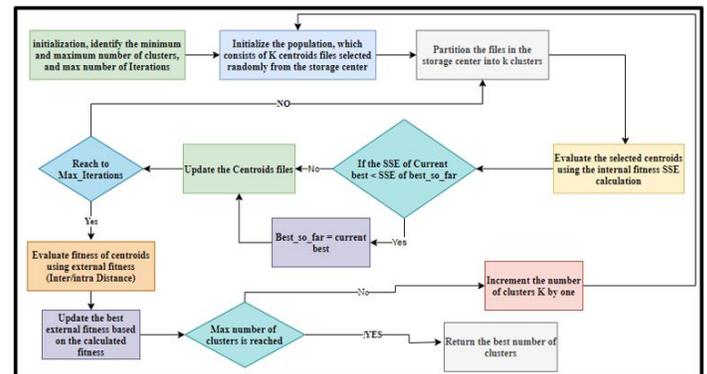


Fig. 3. Main Steps of Genetic-based Dynamic Clustering which used to Divide the Files in the Storage Center of the Cloud Computing.

**Algorithm 1:** Pseudo-code for dynamic clustering algorithm

```

Initialize Define G;
Define N;
Define LP;
Define T;
Define T1;
Define P[size];
Fill T with zeros;
Fill T1 with zeros;
Define I;
Generate permutation lists randomly;
for P = 0; P < L(P); P ++ do
 check the values below for I = 0;
 I < size;
 I ++ do
 Find the row which contains number i in
 permutation list;
 Let R in M(s) = I value;
 for C = I;
 C < Filecount;
 C ++ do
 if R[c] = 1 then
 T[p][c] = i;
 if rowT[p][] contains no zeros then
 Break;
 for I = 1;
 I <= No.Permutation - Lists;
 I ++ do
 Find the row which contains number I in
 permutation list;
 if row T1[p][] contains no zeros then
 setT1[p][1] = i
 if row T1[p][] contains no zeros then
 Break;
Return T;
Return T1;

```

storage center is determined using Euclidean distance, see [19], based on the following equation.

$$D = \sqrt{\sum(C_{jr}, fir)^2} \dots (3)$$

Where D refers to the distance or the similarity between files. Two fitness functions are used through clustering based on genetic, internal fitness, and external fitness. The internal fitness function allows for the selection best centroid data item based on equation (1). The external fitness function has to do with the decision on the proper number of clusters based on the out come of equation (2).

As a summary of the actions and steps of the pre-investigation stage, it starts by dividing the storage center into a set of clusters based on the data types of the files. Afterward, another level of clustering round using genetic-based dynamic clustering, divides the generated clusters from static clustering step into a set of sub-clusters, which produces a set of small clusters for each of the first round's clusters.

The next step in the proposed idea is to start the investigation process by retrieving the cluster that is highly similar to the tampered file based on ranking the clusters according to the similarity achieved through the internal loop of the genetic-based dynamic clustering using equations (1) and (3). Then comes the step of matching and finding the file or files related to the target file. The outcome of the pre-investigation step is to define the clusters that are very similar to the target file. The pre-investigation step can improve the performance and accuracy of the investigation process.

*C. The LSH Algorithm Applied for the Matching and Examination Stage*

Another step in the process of improving the matching and searching for evidence in cloud computing forensics depends on using the Locality Sensitive Hashing (LSH) algorithm. The LSH algorithm consists of three steps. The first step is the shingling step, which is responsible for building the shingle matrix of two dimensions. The second step is the implementation of the Min-hashing algorithm; this step is responsible for the compression of the shingle matrix. The last step is the implementation of the LSH algorithm [24], which is an enhanced matching algorithm that is implemented on the signature matrices generated to identify similar files in the previous step. LSH algorithm can improve the accuracy and efficiency of the investigation process based on reducing the size of the shingle matrix extracted from the original data. Fig. 4 shows the main stages of the LSH algorithm, where the steps begin after retrieving the clusters linked to the tampered file.

The following steps present detailed information on the improved examination and matching steps based on the LSH algorithm.

1) *Shingling Step:* The initial step of the LSH algorithm implementation is the extraction of shingles for each file of the retrieved cluster data, and developing a two-dimensional shingle matrix for all extracted files' shingles. The shingle matrix shown in Table II represents an example of a shingle matrix extracted from the files of the cluster. In the Table, a Shingle matrix ID in a row exists in documents, specified in column, where the flag bit is set.

Table II displays the shingle matrix structure. The shingle matrix building step is the initial step in the LSH algorithm procedure. All documents contents must be transformed into

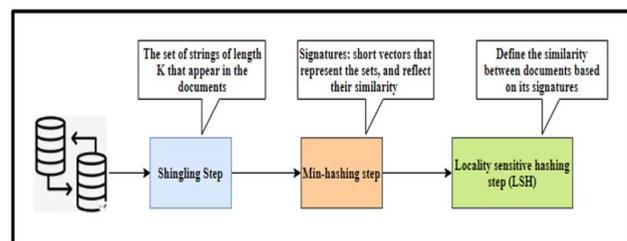


Fig. 4. Main Steps of Applying the LSH Algorithm to Enhance the Investigation Process in Cloud Computing

TABLE II. AN EXAMPLE OF A SHINGLE MATRIX EXTRACTED FOR A SET OF TEXT FILES

| Shingle ID | First Document | Second Document | Third Document |
|------------|----------------|-----------------|----------------|
| Shingle 1  | 1              | 0               | 0              |
| Shingle 2  | 0              | 1               | 0              |
| shingle 3  | 0              | 0               | 1              |
| shingle 4  | 1              | 0               | 0              |
| Shingle 5  | 0              | 1               | 0              |
| Shingle 6  | 0              | 0               | 1              |
| Shingle 7  | 1              | 1               | 0              |
| Shingle 8  | 0              | 0               | 1              |
| Shingle 9  | 1              | 1               | 0              |
| Shingle 10 | 0              | 0               | 1              |
| Shingle 11 | 1              | 1               | 0              |
| Shingle 12 | 0              | 0               | 1              |
| Shingle 13 | 1              | 1               | 1              |
| Shingle 14 | 1              | 1               | 0              |
| Shingle 15 | 0              | 0               | 1              |

shingles that are added into a shingle matrix which represents the files in the cloud storage center.

The process of extracting shingles from documents is called the shingling extraction step where each document is represented by a string of characters. Shingles extraction is the process of creating a set of k-shingles for a document to be any sub string of length k found. As an example of the shingle extraction step, assuming the original document containing a, b, c, d, e, f, g, h, and k is 3 then some possible shingles are a, b, c, b, c, d, c, d, e, d, e, f, e, f, g, and f, g, h. The similarity of documents is measured based on the degree of overlapping between these shingles. Increasing the number of shingles overlapped between two files means that the two files are more similar to each other [11].

The similarity between documents is measured using Jaccard similarity metric [45], which depends on the degree of overlapping between shingles. matrix. The equation used to calculate Jaccard similarity is as follows

$$Sim(File1, File2) = \frac{|File1 \cap File2|}{|File1 \cup File2|}$$

... (4) By increasing the intersection ratio between the two files, the degree of similarity would increase. Based on this, the degree of similarity between the two files will increase when the amount of intersected characters or features between the two files is increased.

2) *The Step of Compression using Min-hashing Algorithm* : The signature matrix which is built through the first step of the pre-investigation stage will be usually very large due to the huge volume of files stored in cloud computing storage space. Accordingly, it is very complicated to handle the shingle matrix for the recovered clusters. Therefore, further processing is needed, which could be very time consuming to use the conventional method for performing the searching and matching steps.

It is necessary to compress the shingle matrix in a specific way while keeping the distance between the files in the original matrix. One of the best solutions is the Min-hashing algorithm, which can compress the shingle matrix into a small matrix called signature matrix "M". The most important feature of this algorithm is that it keeps the similarity of the underlying sets of shingles in the compressed version.

Min-hashing algorithm allows for the generating of a permutation list, which contains random numbers in the range from 1 to the number of shingles. The result of the Min-hashing process is stored in a signature matrix, where its rows are the Min-hashing value, and the column of the signature matrix is the file name. The number of permutation lists determines the accuracy of the signature matrix. Each permutation list produces a row in the signature matrix. Fig. 5 shows an example of the main step of converting the shingle matrix for a set of files into a signature matrix using 4 Min-hashing functions (4 permutation lists). The similarity between doc1 and doc3 in the shingle matrix is very close to the similarity between h(doc1) and h(doc2) in a signature matrix.

The pseudo code shown in Algorithm 1 contains the steps of min-hashing starting from the step of generation of a permutation list and all steps of compression for the shingle matrix until the production of the signature matrix and signature list. For more details on how to calculate the signature matrix out of the permutation lists please refer to [15].

3) *Locality Sensitive Hashing (LSH)*: The last step after creating the signature matrix from the shingle matrix is applying locality-sensitive hashing to the produced signature matrix. LSH step is very helpful in dealing with parts of a file in order to discover criminal acts rather than dealing with the whole file imposed by other techniques. In the matching and examination forensic stage, we suggest to use LSH algorithm to enhance the investigation process based on the following factors:

- 1) The conventional way of matching byte-to-byte takes significantly long time carry out especially on the cloud.
- 2) High degree of accuracy in the matching process can be obtained. This aspect is essential and useful in a cloud computing environment since it contains a huge volume of files.
- 3) The LSH algorithm can handle all cases of the matching process, such as if the file was removed, partially modified, or fully updated.

Locality-sensitive hashing mainly depends on the division of the signature matrix from the shingle matrix into several bands. Specific hash function F(x) is used for each band of a signature matrix. The result of the hashing step maps to a specific bucket in a bucket list. Each column in the signature matrix is hashed multiple times based on splitting the column into band sets, and each band is hashed using a specific, preferably different, hashing function.

The Signature list will also be hashed multiple times based on splitting it into bands and hashing each band using a same hash function that has been used for that band over the signature matrix. The bands with the same data will be hashed and mapped to the same bucket [22]. This means that the files which are shared in the bucket are exactly identical.

The flowchart in Fig. 6 shows the key steps of the LSH algorithm which are used in CBIF as the step of matching began by dividing the signature matrix into a specific number of bands. In order to determine that a set of files match, specific threshold has to be defined to determine which of the candidate set of files have contributed in creating the targeted document. The threshold in CBIF will be changed based on the corresponding results as shown in the flowchart in Fig. 6.

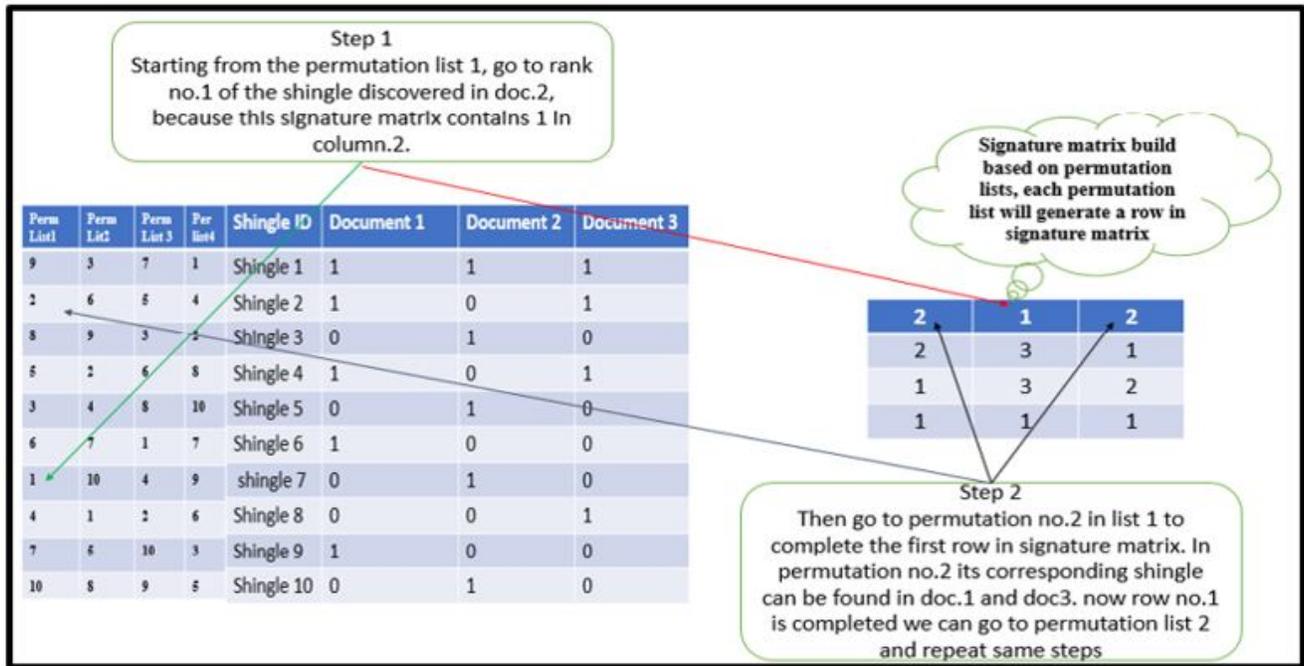


Fig. 5. Main Steps of Building a Signature Matrix from the Shingle Matrix.

```

Algorithm 2: Pseudo code for the main steps of Min-hashing algorithm
Start Define shingle Matrix M(S);
Define shingle List L(x);
Define LP;
Define T;
Define T1;
Define P[size];
Fill T with zeros;
Fill T1 with zeros;
Define I;
Generate permutation lists randomly;
for P = 0; P < P.length; P++ do
 for I = 0; I < size; I++ do
 Find the row which contains number i in permutation list;
 Let R in M(S) = I;
 for C = I; C < File - Count; C++ do
 if R[C] = 1 then
 T[P][C] = i
 rowT[P][C] contains no - zeros Break;
for P = 1; P <= P.length; P++ do
 for I = 1; I <= LP; I++ do
 Find the row which contains number I in permutation list if rowT1[P][I] contains no zeros then
 T1[P][I] = i;
 rowT1[P][I] contains no zeros Break;
Return T, T1;

```

If there are no two files shared in a number of buckets based on the given threshold, the threshold's value will be reduced until it reaches zero. Therefore, it can be concluded that no file matches fully or partially the targeted file. Fig. 7 shows an example of how to apply LSH algorithm in the matching system to identify the set of matching documents.

#### IV. EXPERIMENTS IMPLEMENTATION AND RESULTS DISCUSSION

Two main experiments have been implemented throughout this research paper. The first one has to do with the implementation and evaluation of the clustering algorithm which makes use of CBIF. The second experiment has to do with the implementation and evaluation of LSH algorithm. All the experiments were implemented with Java Programming language using Net-Beans 8.1 on a PC with 64-bit Windows 10 operating system, and an Intel core i7-6500u CPU with 32GB of RAM.

##### A. Testing and Evaluation of Clustering Algorithm

To define the performance of the genetic-based dynamic clustering algorithm used through the second level of clustering, an implementation of the algorithm over ten, real and artificial data-sets obtained from UCI [10] was performed. The data-sets used in this experiment have various numbers of features ranging between '3' to '13' features. It contains various numbers of clusters ranging from '2' to '6' clusters. A set of experiments have been designed to measure the accuracy and performance of GBDCA before being used in CBIF. Each experiment has been applied over the data-sets mentioned in Table III. GBDCA has been compared with K-means clustering algorithm which we have been also implemented to measure its accuracy and performance.

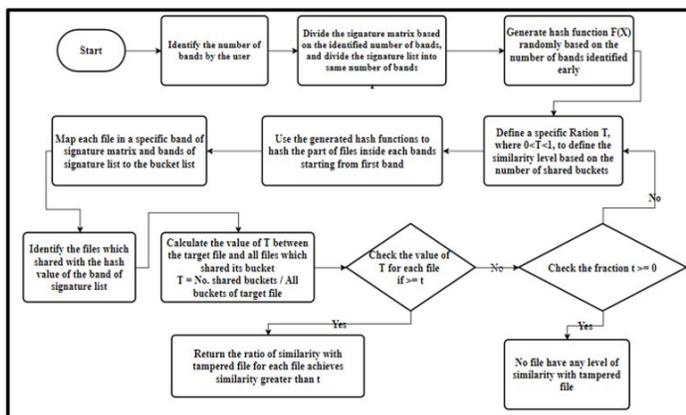


Fig. 6. Main Steps of Locality Sensitive Hashing Algorithm for Matching.

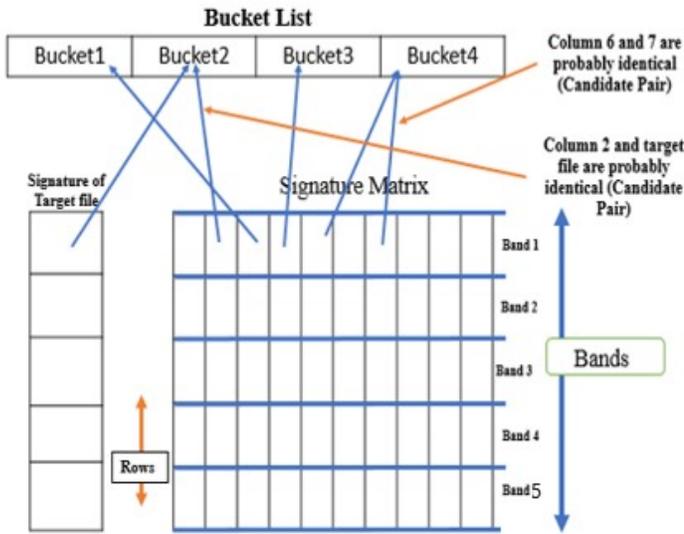


Fig. 7. A Locality-sensitive Hashing Algorithm (LSH) and How the Matching Process is Done based on the Map for One Band.

TABLE III. EVALUATION RESULTS OF THE PROPOSED DYNAMIC CLUSTERING ALGORITHM (GBDCA).

| Dataset Name | Original No of Clusters | Avg. No of clusters GBDC | Avg. No of clusters CRO | Avg No.clusters Kmeans |
|--------------|-------------------------|--------------------------|-------------------------|------------------------|
| E.coli DS    | 5                       | 4.9                      | 5.1                     | 4.7                    |
| IRIS DS      | 3                       | 3.43                     | 3.63                    | 4.23                   |
| Seed DS      | 3                       | 3.367                    | 3.53                    | 4.134                  |
| Balance DS   | 3                       | 3.23                     | 3.36                    | 4.1                    |
| Blood DS     | 2                       | 2.5                      | 2.6                     | 2.934                  |
| Wine DS      | 3                       | 2.8                      | 2.53                    | 3.567                  |
| Hepatitis DS | 2                       | 2.76                     | 2.83                    | 3.067                  |
| Vertebral DS | 2                       | 2.73                     | 2.93                    | 3.43                   |
| BC DS        | 2                       | 2.36                     | 2.57                    | 3.034                  |
| Glass DS     | 6                       | 5.033                    | 5.63                    | 4.863                  |

Table III represents the number of clusters resulted from running each of the two clustering algorithms. The number of clusters achieved by each algorithm is compared with the original number of clusters for each data-set, and represented as the Error Rate. The results shown in Table III is the average value of ten experiment runs applied for each data-set. The main objective of these experiments and the comparison between GBDCA and K-means algorithms were to prove that GBDCA is the better than K-means and can be used in the investigation process to enhance the forensics process of digital devices. The genetic based dynamic clustering algorithm depends on using the Inter/Intra ratio, as an external fitness function.

$$POE = \frac{ABS(OriginalNo - AchievedNo)}{Originalnumber} * 100$$

... (5) In the equation above, the POE refers to the percentage of error, ABS refers to the absolute value, which is calculated for the difference between the original number (OriginalNo.) of clusters for each algorithm and the average number of clusters achieved by the experiments applied on a specific data-set (AchievedNo.). Table IV shows the error rate for each competitive algorithm for the different data-set.

TABLE IV. ERRORS PERCENTAGE FOR EACH COMPETITIVE CLUSTERING ALGORITHM BASED ON THE RESULTS ILLUSTRATED IN TABLE 2

| Dataset Description | Genetic Error Rate | CRO Error Rate | Kmeand Error Rate | Min Error Rate |
|---------------------|--------------------|----------------|-------------------|----------------|
| E. coli DS          | 10%                | 14.67%         | 44.67%            | 10%-GA         |
| IRIS DS             | 14.44%             | 21.11%         | 41.1%             | 14.44%-GA      |
| Seed DS             | 12.22%             | 18%            | 37.78%            | 12.22%-GA      |
| Balance DS          | 7.78%              | 12%            | 36.6%             | 7.78%-GA       |
| Blood DS            | 25%                | 30%            | 46.67%            | 25%-GA         |
| Wine DS             | 6.67%              | 16%            | 18.89%            | 6.67%-GA       |
| Hepatitis DS        | 38.33%             | 42%            | 53.33%            | 38.33%-GA      |
| Vertebral DS        | 36.67              | 47             | 53.33%            | 36.67%-GA      |
| Breast Cancer       | 18.33%             | 28%            | 51.7%             | 18.33%-GA      |
| Glass DS            | 16.11%             | 6%             | 18.9%             | 6%-CRO         |
| Average Results     | 19%                | 23%            | 40%               | 19%-GA         |

Table IV allows us to determine which is the best algorithm to be used through the step of dynamic clustering in the proposed hierarchical clustering algorithm. The average percentage of error rate achieved by the dynamic clustering-based genetic algorithm is 19%, and the percentage of error achieved by the dynamic conventional k-means algorithm is 40%. The results shown in Table 3 can be interpreted as a recommendation to use a genetic-based dynamic clustering algorithm in our solution, given the low percentage rate in comparison with the K-means algorithm.

### B. Experimental Implementation for the LSH Algorithm

Several tests have been performed to determine the accuracy and reliability of CBIF and different data set sizes have been used. Furthermore, various structures for the tampered file were designed and used to test the proposed solution through a set of experiments. The data set used in our experiments is the Reuters data set consisting of 21000 files [46]. The experiments are divided into four main categories based on the structure of the tampered file used. Each category of experiments consists of 10 runs and the average value was calculated. In each run, the variables that controls the structure of tampered file were different. The details about the structures of tampered file will be shown in detail just before discussing the results of that experiment. The Reuters files were divided into sets and the series of experiments were applied to each of these sets; the size of these sets will be addressed below. The matching accuracy based on equation (5) has also been calculated.

The data set used in these experiments is divided into clusters based on the dynamic clustering algorithm decision proposed in the hierarchical clustering step. In most cases, the number of clusters achieved was 6 clusters. The files were distributed over the clusters based on the level of similarity between cluster centroid files and all other files.

1) *Tampered File Description:* The tampered file used in the experiments to evaluate the accuracy and efficacy of the CBIF has many structures, as follows.

- Structure-One: The first structure for the tampered file

consists of three random files from three random clusters. File "A", from a random cluster "A", is involved with 10% of the tampered file content. Random file "B", from another random cluster "B", is involved with 20% of the tampered file content. Finally, 70% of the tampered file creation process is from random file "C", which belongs to yet another cluster "C". The files and clusters "A", "B", and "C" are random in that they may vary from experiment to experiment. This also applies for the next structures. see Fig. 8A.

- Structure-Two: The second structure for the tampered file consists of three random files from three random clusters with different participation rates with 20%, 30%, and 50% for random files A, B, and C from random clusters A, B, and c respectively. Fig. 8B shows the percentage of participation for each file in the tampered file content data
- Structure-Three: Another structure for the tampered file was used that consists of two random files from two random clusters. Random file "A", from random cluster "A", is involved with 40% of the tampered file content, while random file "B", from random cluster "B", is involved with 60% of the tampered file content. Third part of Fig. 8C shows the percentage of participation for each file in the tampered file content.
- Structure-Four: The last structure of the tampered file used consists of two random files "A" and "B" from two random clusters "A" and "B" where each file is involved with 50% of the tampered file as shown in Fig. 8D.

2) *Experimental Implementation for LSH Based Framework Using Different Tampered File's Structures* : Multiple experiments were conducted based on the above mentioned tampered file structures. Each experiment was replicated sixty times with random file selected randomly from a different cluster each time. The AMA is the average results for '60' of the repeated experiments. Each experiment has different variable values for the structure of the tampered file. The purpose of repeating the experiments '60' times is to test CBIF's performance and reliability in order to deal with all cases of file creation being tampered with. The purpose of repeating the experiments is to test the CBIF's ability to recognize the original files in all cases of manipulated files.

For example, in the experiment defined for Structure-One of the tampered file structure, Cluster A can be randomly selected as Cluster '2', and File X can be randomly selected to be File '5', meaning that the tampered file has 10% of its content taken randomly from File number '5' in Cluster '2'. Cluster B can be cluster '4' and File Y can be File '1', this means that the randomly selected File '1' in Cluster '4' has participated with 20% of that tampered file content. Finally, Cluster C can be cluster '5' and File Z can be File '2', meaning that from Cluster '5' we have randomly selected File '2', which contributes to 70% of that same tampered file's content. Fig.9 consists of 4 main sub-figures A,B,C, and D, whereas each sub-figure of Fig. 9 is specified for an experiment based on a specific tampered file structure.

The Average Matching Accuracy (AMA) is a metric that we have used to measure the accuracy of the matching process

and is calculated based on the percentage of matching between the tampered file and the original files which participated in the tampered file content. The equation which was used to calculate AMA between the tampered files and original files is equation (6). In this equation the Avg-Matching(Bx, ..., By) refers to the average results of matching between the tampered file and all clusters participating with a certain percentage of the tampered file's construction. As an example, if equation (6) was used to calculate the average accuracy for a cluster involved with 50% of the tampered file content, then in this case, the AvgMatching(F.1, F.2, ..., F.n) is the average result of matching between the tampered file and the "n" files F.1, ..., F.n that are involved with 50% of the tampered file content. Here "n" is the number of runs the experiment has been repeated which is in our case equals 60.

$$AMA = 100 - \frac{ABS(AvgMatching(Bx, \dots, By) * 100 - Actualmatchingpercentage)}{ActualmatchingPercentage} * 100 \dots (6)$$

The results in Fig. 9 show how accurate CBIF is for the matching process, and presents how the AMA for the results of matching is close to the actual ratio of participation in all matching results.

Details of each result shown in Fig. 9 will be thoroughly explored next before comparing CBIF with other rival algorithms. The purpose of this is to reduce the complexity of discussing this comparison into two simpler phase. In the first phase we discuss the results obtained for CBIF and in the second phase we compare these results with results obtained for other protocols.

- **First Experiment**

The tampered file used through the first experiment was built according to Structure-One of the tampered file structures shown in Fig. 9A, which consists of three files.

Fig. 9-A consists of three columns where each column shows the average matching ratio between the tampered file and the original files in a specific cluster based on the percentage of participation in the tampered file content. The figure show the participation results of three clusters, thus we have three columns, since the participation ratios of other clusters were negligible and thus are not shown in the figure.

The average accuracy of matching between the tampered files and original files which were involved in 10% of tampered file contents is 76.78%, the average accuracy of matching with the files involved in 20% of tampered file contents is 94.44%, and the average result of matching with the files participating in 70% of the content is 87.71%.

The conclusion that can be drawn from Fig. 9A is that CBIF's has the ability to deal with the case when the target file in the investigation process was mixed with other files. This indicates that the file that the investigator was searching for and modified at a certain rate was found at a high accuracy rate.

- **Second Experiment**

Another experiment was conducted based on the second structure of the tampered file that was referred to in Fig. 9B. This figure consists of three columns, each column is specified for the AMA between

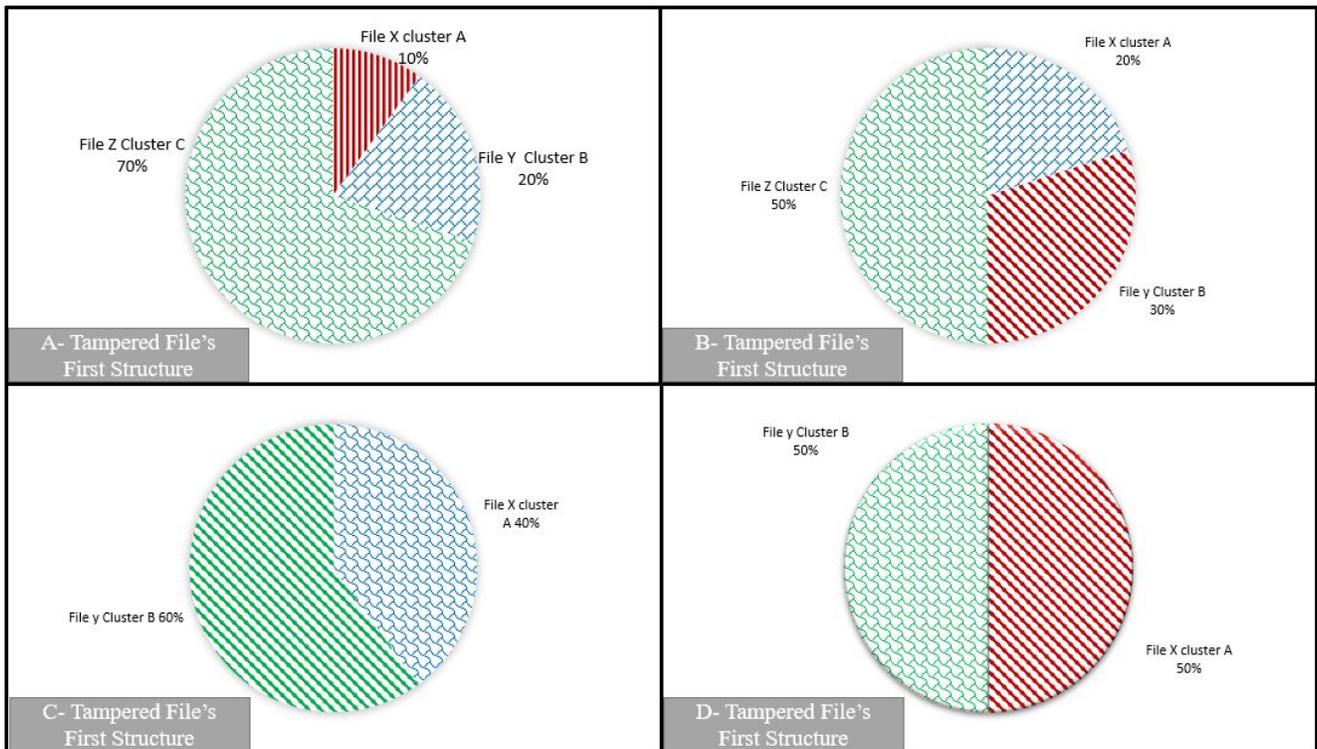


Fig. 8. Key Structures of Tampered File Content which were used by Experimental Implementation.

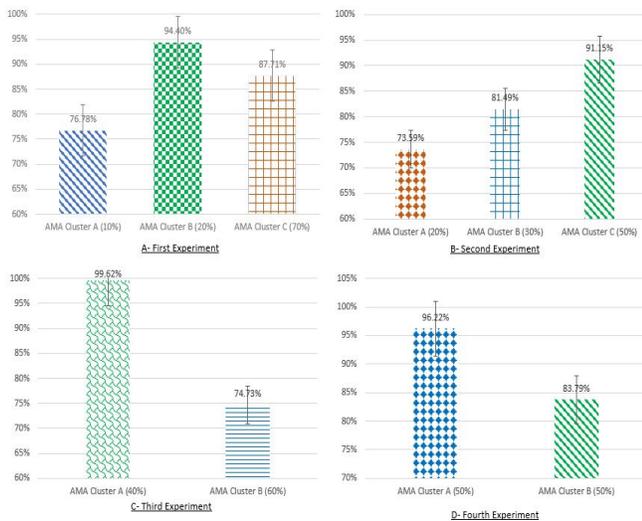


Fig. 9. Results for Set of Experiments based on Different Structures for Tampered Files.

tampered file and each of the original files which were involved in the creation process of the tampered file.

The files that participated in the tampered file's content creation achieves a high level of similarity if compared with other files, such that the accuracy of matching that was shown in Fig.9B depends on

the values of actual matching results. The conclusion that can be drawn on the basis of the results shown in Fig.9B is that the proposed CBIF succeeds in retrieving the right files with a higher matching level than the other files in the storage space.

- Third Experiment The third experiments was applied based on the third structure of the tampered file mentioned earlier in Fig. 9C. This figure consists of two columns, each column represents the average accuracy of matching between the tampered files and original files which were involved in the construction process of the tampered file. The results shown in Fig. 9C represents the extent of accuracy and reliability of CBIF to deal with the third case of tampered file content mentioned earlier. This gives a strong argument for the benefits and importance of using the proposed CBIF.
- Fourth Experiment. The last experiment was applied based on the structure of the tampered file mentioned in Fig. 9D. The results shown in Fig. 9D illustrate the AMA for this experiment which consists of two bars, each bar shows the average accuracy of matching with each original file involved in creating the tampered file. The results shown in Fig. 9D show the accuracy of CBIF in retrieving the correct files which actually participated in the contents of the manipulated file.

3) Comparisons between the Proposed Solution and Related Cluster-based Solutions: Two of the related solutions have been implemented and compared with the proposed CBIF.

The purpose of this comparison is to demonstrate matching accuracy level of CBIF relative other similar techniques. The first competitive solution relies on using the traditional k-means clustering algorithm [10]. The idea of the solution revolves around applying the k-means clustering algorithm in the investigation process to divide the documents in the storage center into groups and to focus on the group which is most similar to the target file. The second competitive solution is to use a k-medoid clustering algorithm to enhance the investigation process with respect to accuracy and performance [17]. Both solutions have been implemented and used with the data-set which was used to test CBIF. The structure of the tampered file which was used through the experiments of testing the competitive solutions is the same as the structure which mentioned earlier to test the proposed CBIF. All of experiments have been run 60 times. The average accuracy of matching for all experiments have been applied using equation 6 . The results for the experiments are as follows:

- Comparison between CBIF, a K-means based solution, and a k-medoid based solution using the first structure of the tampered file

Fig. 10A shows the AMA for the three solutions based on structure-One of the tampered file. The figure consists of three groups of bars, each group contains three bars. Each bar refers to the average accuracy of matching between the tampered file and other files, exist in a specific cluster, using one of the three compared solutions. The first group which represents the AMA between the tampered file and a clustered file which participated in 10% of the tampered file content. CBIF has achieved higher AMA than the results obtained by the other solutions for the first tampered file structure.

The AMA of the file which participated with 70% of tampered file content is greater in CBIF, close to 88%, compared to the other solutions, almost 75% and 65% for k-means and k-medoid, respectively. This demonstrates the ability of CBIF in improving the matching accuracy which is better than similar solutions. This helps in finding the correct file that actually participated in the contents of the tampered file.

Another important enhancement of CBIF is related to the rank of the original files which participated in the tampered file's content. The files that were already involved in creating the file that was tampered with, in most experiments, have the highest degree of similarity with the tampered file. On the other hand, only in 72% of the total experiments did the original files that participated in the composition of the manipulated file obtain the highest similarity with the tampered file, which means that the original files did not get the highest similarity in a 28% of experiments.

- Comparison based on the second structure of the tampered file  
Other experiments have been designed and implemented for the related solutions to show how CBIF solution performs. AMA has been calculated for the 60 runs of experiment for each solution. The summary

of these experiments can be shown in Fig. 10B. Results shown in Fig. 10B represent the AMA of the competitive solutions to retrieve the original participated files. CBIF achieved an accuracy level that is, also in this experiment, higher than other two competitive solutions for all participated files, as can be seen in the results shown in group three of Fig. 10B.

The AMA achieved by CBIF is 91.15%, whereas the average accuracy achieved by the k-means based solution is close to 75%, and the average accuracy of matching achieved by the k-medoid based solution is close to 67%. This indicates the gap in the performance between CBIF solution and the other related solutions. This point is important to encourage the prefer the usage of the CBIF in investigation process over the other rival solutions.

Another significant improvement accomplished by CBIF is the ranking of files based on the degree of similarity with the tampered file. File ranking here indicate the ability of the solution to rank the actual original files that have participated in the tampered file as the top ones. For example assume that file A and B have participated in creating a tampered file T with a certain percentage from each one. A solution that gives the highest percentages to files A and B, regardless of the accuracy results for each one, is assumed to reach 100% ranking accuracy. While a solution that gives a percentage for another file, e.g. file C, that is higher than A or B gets a lower ranking accuracy.

In the experiments that are based on the Structure-One of the tampered file, using a solution based on k-means and a solution based on k-medoid, the files involved with the tampered file content do not reach the highest degree of similarity with the tampered file in some of the experiments. In the solution that depends on the k-medoid clustering algorithm, the ranking accuracy was close to 77% while in the solution that depends on the conventional k-means clustering algorithm, 80% was the ranking accuracy. CBIF has achieved 95% to 100% ranking accuracy in all of the experiments and for all tampered file structures.

- Comparison based on Structure-Three of the tampered file.  
Third part of Fig. 10C shows the results for the experimental implementation of competitive solutions based on Structure-Three of the tampered file mentioned earlier. For the 60 experimental runs, the average accuracy was calculated. The average level of accuracy obtained by CBIF was higher than the average accuracy achieved by the competitive solutions in the matching process with the files that participated in 40% of the tampered file contents. Although the difference is less than the case with 40% participation, CBIF achieved higher accuracy also for the 60% participation results. Additionally, for the ranking accuracy, in the k-means based solution 86% of all experiments. In the k-medoid based solution is 84% of all experiments.
- Comparison between the CBIF and other competitive solutions based on the Structure-Four of the tampered file

The results shown in last part of Fig. 10D represent the average accuracy of matching for the three competitive solutions to enhance the investigation process. CBIF achieves higher AMA in both clusters, which has participated with 50% of the tampered file, than the other compared solutions. Moreover, k-medoid based solution, shows ranking accuracy as 86% while solutions based on k-means results in 90% ranking accuracy which is much less than what CBIF has achieved.

To conclude, from the section of comparisons between the proposed CBIF and other rival solutions, it can be inferred that the proposed solution improves the average matching accuracy for the investigation process in all cases and for all of the files structures involved. This point is important in the forensics process, as many of the crimes are hidden based on mixing the original file with other files. Moreover, the point of enhancement in the accuracy of matching is what all frameworks and solutions are looking for to achieve; the results shown in the above figure indicate a noticeable improvement on average matching accuracy in most cases. This is supported by the results of ranking accuracy which prove that CBIF outperforms K-means and K-mediod algorithms in all the conducted experiments.

4) *Comparison between CBIF and Related Solution on the Basis of Execution Time:* Another comparison is between the competitive solutions based on the execution time for each algorithm. The comparison is based on the size of the data-set used. The execution time was measured for applying the three solutions for the 2000 file, 4000 file, 8000 file, 16000 file, and 21000 file data-sets. The results for the execution time are shown in Table V. The results in Table V refers to the time needed for the investigation to find a file similar to the target file. k-medoid clustering algorithm needs less time than k-means based investigation process which in turn takes less than proposed CBIF. This slight extra is justified by the higher average matching accuracy achieved by CBIF.

TABLE V. RESULTS FOR THE EVALUATION PROCESS OF THE PROPOSED DYNAMIC CLUSTERING ALGORITHM

| No.Files in a data-set | Execution time for CBIF in seconds | Execution time For k-means based investigation in seconds | Execution time for k-medoid in seconds |
|------------------------|------------------------------------|-----------------------------------------------------------|----------------------------------------|
| 1000                   | 189.2                              | 150.388                                                   | 135.9                                  |
| 2000                   | 238.7                              | 215.9                                                     | 182.6                                  |
| 4000                   | 1972.15                            | 1213                                                      | 600.356                                |
| 8000                   | 4370.16                            | 4222.11                                                   | 3446.462                               |
| 16000                  | 12325.2                            | 11863.5                                                   | 10695                                  |
| 21000                  | 18735.5                            | 17962                                                     | 16818.5                                |

## V. CONCLUSION

Many frameworks and solutions have been proposed in the field of cloud computing forensics. Each of these frameworks have specific drawbacks and weak points in certain stages of the investigation process. In the paper, CBIF is proposed to focused on a specific issue of cloud computing forensics; namely, the performance and accuracy of the investigation process. CBIF adds a new stage called the pre-investigation stage, which is responsible for filtering and grouping the evidence and files in the cloud storage center into a set of groups based on using a hierarchical clustering approach. This stage enhances the average accuracy of the matching process by dividing the storage center into sets of groups.

The results achieved from the experimental implementation shows how the proposed CBIF enhances the average accuracy when compared with the related solutions, the k-means based solution and the k-medoid based solution. CBIF achieved higher accuracy matching levels and ranking accuracy than the competing solutions in most of the experiments.

The proposed CBIF solution can be enhanced in the future by enhancing the LSH step of current solution by changing the technique of selecting band size.

## REFERENCES

- [1] M Edington Alex and R Kishore. Forensics framework for cloud computing. *Computers & Electrical Engineering*, 60:193–205, 2017.
- [2] Wesam Almobaideen and Muhyidean Altaramneh. Fog computing: survey on decoy information technology. *International Journal of Security and Networks*, 15(2):111–121, 2020.
- [3] Wesam Almobaideen and Ola Malkawi. Application based caching in fog computing to improve quality of service. In *2020 Fifth International Conference on Fog and Mobile Edge Computing (FMEC)*, pages 20–27. IEEE, 2020.
- [4] Arushi Arora, Sumit Kumar Yadav, and Kavita Sharma. Denial-of-service (dos) attack and botnet: Network analysis, research tactics, and mitigation. In *Handbook of Research on Network Forensics and Analysis Techniques*, pages 117–141. IGI Global, 2018.
- [5] Mustafa Aydin and Jeremy Jacob. A comparison of major issues for the development of forensics in cloud computing. In *8th International Conference for Internet Technology and Secured Transactions (ICITST-2013)*, pages 77–82. IEEE, 2013.
- [6] Leonardo Babun, Amit Kumar Sikder, Abbas Acar, and A Selcuk Uluagac. Iotdots: A digital forensics framework for smart environments. *arXiv preprint arXiv:1809.00745*, 2018.
- [7] Stacey O Baror, Hein S Venter, and Richard Adeyemi. A natural human language framework for digital forensic readiness in the public cloud. *Australian Journal of Forensic Sciences*, 53(5):566–591, 2021.
- [8] Venansius Baryamureeba and Florence Tushabe. The enhanced digital investigation process model. *Digital Investigation*, 2004.

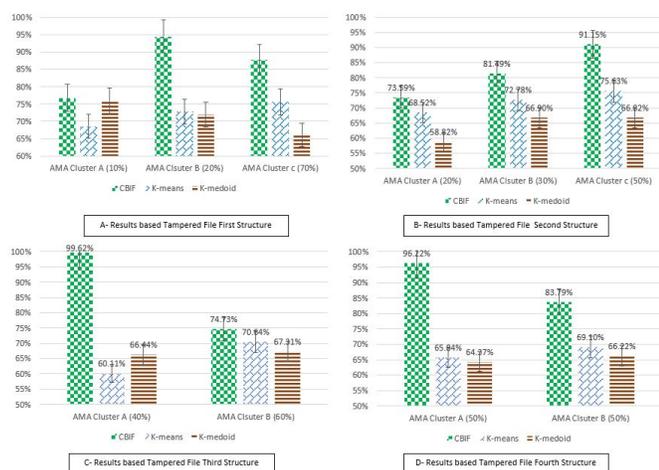


Fig. 10. Results for All Competitive Algorithms based on Different Structures for Tampered File Content.

- [9] Ulrich Bayer, Paolo Milani Comporetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. Scalable, behavior-based malware clustering. In *NDSS*, volume 9, pages 8–11. Citeseer, 2009.
- [10] Catherine Blake. Uci repository of machine learning databases. <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998.
- [11] Andrei Z Broder. Identifying and filtering near-duplicate documents. In *Annual Symposium on Combinatorial Pattern Matching*, pages 1–10. Springer, 2000.
- [12] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
- [13] Mumin Cebe, Enes Erdin, Kemal Akkaya, Hidayet Aksu, and Selcuk Uluagac. Block4forensic: An integrated lightweight blockchain framework for forensics applications of connected vehicles. *IEEE Communications Magazine*, 56(10):50–57, 2018.
- [14] Yu-Jia Chen and Li-Chun Wang. A security framework of group location-based mobile applications in cloud computing. In *2011 40th International Conference on Parallel Processing Workshops*, pages 184–190. IEEE, 2011.
- [15] Ondrej Chum, Michal Perd'och, and Jiri Matas. Geometric min-hashing: Finding a (thick) needle in a haystack. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24. IEEE, 2009.
- [16] Jasmin Čosić, Zoran Čosić, and Miroslav Baća. An ontological approach to study and manage digital chain of custody of digital evidence. *Journal of Information and Organizational Sciences*, 35(1):1–13, 2011.
- [17] Luis Filipe da Cruz Nassif and Eduardo Raul Hruschka. Document clustering for forensic analysis: An approach for improving computer inspection. *IEEE transactions on information forensics and security*, 8(1):46–54, 2012.
- [18] Larry Daniel. *Digital forensics for legal professionals: understanding digital evidence from the warrant to the courtroom*. Elsevier, 2011.
- [19] Per-Erik Danielsson. Euclidean distance mapping. *Computer Graphics and image processing*, 14(3):227–248, 1980.
- [20] Sergio Decherchi, Simone Tacconi, Judith Redi, Alessio Leoncini, Fabio Sangiacomo, and Rodolfo Zunino. Text clustering for digital forensics analysis. In *Computational Intelligence in Security for Information Systems*, pages 29–36. Springer, 2009.
- [21] Athanasios Dimitriadis, Nenad Ivezic, Boonserm Kulvatunyou, and Ioannis Mavridis. D4i - digital forensics framework for reviewing and investigating cyber attacks. *Array*, 5:100015, 2020.
- [22] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999.
- [23] David E Goldberg. Genetic algorithms in search. *Optimization, and Machine Learning*, 1989.
- [24] Christopher Hargreaves and Jonathan Patterson. An automated timeline reconstruction approach for digital forensic investigations. *Digital Investigation*, 9:S69–S79, 2012.
- [25] Ben Hitchcock, Nhien-An Le-Khac, and Mark Scanlon. Tiered forensic methodology model for digital field triage by non-digital evidence specialists. *Digital investigation*, 16:S75–S85, 2016.
- [26] Michael Hogan, Fang Liu, Annie Sokol, and Jin Tong. Nist cloud computing standards roadmap. *NIST Special Publication*, 35:6–11, 2011.
- [27] John H Holland. Genetic algorithms. *Scientific american*, 267(1):66–73, 1992.
- [28] Liu Jiang, Guiyan Tian, and Shidong Zhu. Design and implementation of network forensic system based on intrusion detection analysis. In *2012 International Conference on Control Engineering and Communication Technology*, pages 689–692. IEEE, 2012.
- [29] Karen Kent, Suzanne Chevalier, Tim Grance, and Hung Dang. Guide to integrating forensic techniques into incident response. *NIST Special Publication*, 10(14):800–86, 2006.
- [30] Mohammad Khanafsa, Ola Surakhi, and Sami Sarhan. A parallel face detection method using genetic & cro algorithms on multi-core platform. In *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, pages 1–6. IEEE, 2019.
- [31] Mohammed Khanafseh, Mohammad Qatawneh, and Wesam Almobaideen. A survey of various frameworks and solutions in all branches of digital forensics with a focus on cloud forensics. *Int. J. Adv. Comput. Sci. Appl*, 10(8):610–629, 2019.
- [32] Manish Kumar, M Hanumanthappa, and TV Suresh Kumar. Network intrusion forensic analysis using intrusion detection system. *Int. J. Comp. Tech. Appl*, 2(3):612–618, 2011.
- [33] Chanjin Lee and Mokdong Chung. Digital forensic for location information using hierarchical clustering and k-means algorithm. *Multimedia Society Journal*, 19(1):30–40, 2016.
- [34] Sheik Khadar Ahmad Manoj, D Lalitha Bhaskari, et al. Cloud forensics-a framework for investigating cyber attacks in cloud environment. *Procedia Computer Science*, 85:149–154, 2016.
- [35] Raffael Marty. Cloud application logging for forensics. In *proceedings of the 2011 ACM Symposium on Applied Computing*, pages 178–184, 2011.
- [36] Matthias Neuschwandtner, Paolo Milani Comporetti, Gregoire Jacob, and Christopher Kruegel. Forecast: skimming off the malware cream. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 11–20, 2011.
- [37] Liwen Peng, Xiaolin Zhu, and Peng Zhang. A framework for mobile forensics based on clustering of big data. In *2021 IEEE 4th International Conference on Electronics Technology (ICET)*, pages 1300–1303, 2021.
- [38] Mohammad Qatawneh, Ahmad Alamoush, and Ja'far Alqatawna. Section based hex-cell routing algorithm (sbhcr). *International Journal of Computer Networks & Communications*, 7(1):167, 2015.
- [39] Mohammad Qatawneh et al. Multilayer hex-cells: A new class of hex-cell interconnection networks for massively parallel systems. *IJCNS*, 4(11):704–708, 2011.
- [40] Mark Reith, Clint Carr, and Gregg Gunsch. An examination of digital forensic models. *International Journal of Digital Evidence*, 1(3):1–12, 2002.
- [41] M Rogers. *Dcsa: a practical approach to digital crime scene analysis*. West Lafayette, Purdue University, 2006.
- [42] Keyun Ruan, Joe Carthy, Tahar Kechadi, and Ibrahim Baggili. Cloud forensics definitions and critical criteria for cloud forensic capability: An overview of survey results. *Digital Investigation*, 10(1):34–43, 2013.
- [43] Huda K Saadeh, Wesam Almobaideen, and Khair Eddin Sabri. Ppustman: Privacy-aware publish/subscribe iot mvc architecture using information centric networking. *Modern Applied Science*, 12(5):128, 2018.
- [44] Sarah Shukri, Hossam Faris, Ibrahim Aljarah, Seyedali Mirjalili, and Ajith Abraham. Evolutionary static and dynamic clustering algorithms based on multi-verse optimizer. *Engineering Applications of Artificial Intelligence*, 72:54–66, 2018.
- [45] Jatsada Singthongchai and Suphakit Niwattanakul. A method for measuring keywords similarity by applying jaccard's, n-gram and vector space. *Lecture Notes on Information Theory*, 1(4), 2013.
- [46] Kilian Stoffel, Paul Cotofrei, and Dong Han. Fuzzy methods for forensic data analysis. In *2010 International Conference of Soft Computing and Pattern Recognition*, pages 23–28. IEEE, 2010.
- [47] Akash A Thakar, Kapil Kumar, and Baldev Patel. Next generation digital forensic investigation model (ngdfim)-enhanced, time reducing and comprehensive framework. In *Journal of Physics: Conference Series*, volume 1767, page 012054. IOP Publishing, 2021.
- [48] Sebastiaan Von Solms, Cecil Louwrens, Colette Reekie, and Talania Grobler. A control framework for digital forensics. In *IFIP International Conference on Digital Forensics*, pages 343–355. Springer, 2006.
- [49] Wei Wang and Thomas E Daniels. A graph based approach toward network forensics analysis. *ACM Transactions on Information and System Security (TISSEC)*, 12(1):1–33, 2008.
- [50] Shams Zawoad, Amit Kumar Dutta, and Ragib Hasan. Seclaas: secure logging-as-a-service for cloud forensics. In *Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security*, pages 219–230, 2013.
- [51] Tanveer Zia, Peng Liu, and Weili Han. Application-specific digital forensics investigative model in internet of things (iot). In *Proceedings of the 12th International Conference on Availability, Reliability and Security*, pages 1–7, 2017.

# Towards a Low-Cost FPGA Micro-Server for Big Data Processing

Mohamed Abouzahir<sup>1</sup>

Ecole Supérieure de Technologie de Salé  
Laboratoire LASTIMI  
Université Mohammed V de Rabat

Khalifa Elmansouri<sup>2</sup>

Institut Supérieur des Sciences de la Santé (ISSS)  
Laboratoire des Sciences et Techniques de la Santé  
Université Hassan 1er, Settat

Rachid Latif<sup>3</sup>

Ecole Nationale des Sciences Appliquées d'Agadir  
Laboratoire d'Ingénierie des Systèmes et Technologies de l'Information (LISTI)  
Université Ibn Zohr, Agadir

Mustapha Ramzi<sup>4</sup>

Ecole Supérieure de Technologie de Salé  
Laboratoire LASTIMI  
Université Mohammed V de Rabat

**Abstract**—The development of big data in the era of data explosion and the growing demand for micro-servers in place of traditional servers to adapt to lightweight tasks in recent years has put into question how to integrate and make use of these two important domains. During the same era, CPU performance growth has reached a certain maturity. In order to surpass these issues and to reach high performances computing, a new trend now is to use multiple processing units or heterogeneous components in micro-servers to reduce computational complexity. The implementation of Big Data processing algorithms using embedded heterogeneous architectures rises a new challenges due to constraints of the used architecture-based system on chip which require a special attention and imposed new demands to our works. In this article, we focus on using embedded FPGA accelerator to give a solution to this problem. Precisely, we will attempt to prototype a micro-server for the processing of big data on FPGA and compare its performances with a high-end GPGPU using existing benchmarks. The implementation on the FPGA is done using a High-Level Synthesis based-OpenCL (HLS) instead of the traditional description language. The obtained results shows that FPGA is an interesting alternative and can be a promising platform to design a micro-server when it comes to process a hug amount of data, in particular with the emerging technologies for FPGA programming using HLS approach and by adopting the OpenCL optimization strategies.

**Keywords**—Arria 10 FPGA (Field Programmable Gate Arrays); GPGPU (General Purpose Graphics Processing Unit) ; big data; parallel computing; (HLS) High-Level Synthesis

## I. INTRODUCTION

BIG data, produced from online transactions, emails, videos, audios, picture, posts, search interrogation, medical records, social networking interface and science applications, has become one of the most important domain in the information technology industry [1]. According to IBM (2015), around 2.5 nonillion (10 to the power of 30) bytes of data is created every day. The overwhelmingly large amount of data leads to challenges including storage, analysis and fetch. To define the properties of big data, Doug Laney (2001) has proposed the 3Vs (Volume, Variety, Velocity) [2]. In current hypercompetitive industry environment, several challenges related to big data processing are imposed on the companies, which include to meet with the need for speed, to understand the data, to address

the data quality, to display significant results and to distinguish the outliers. There is growing interest in FPGA (Field Programmable Gate Array) as a solution. FPGA has always been considered as the generation of integrated circuit that could replace ASIC (application specific integrated circuit). Able to be fully reconfigured by user, an FPGA is commonly claimed higher performance, shorter time to market, lower cost, high reliability, less needs for long-term maintenance. Being reprogrammable results is the main difference between an FPGA and an x86 processor: the FPGA does not waste compute cycles doing unnecessary processing. In other words, FPGA excels for doing one simple and repetitive task like pattern matching. This great advantage of FPGA has catalyzed investigations on its potential usage in big data processing. One of the most important assets of FPGA is its exceptional ability in the computation of finegrained tasks in clock-cycle basis. This ability is extremely interesting especially regarding Big Data management. However, this high-potential acceleration constrains drastically the possibilities of implementation. After having chosen the FPGA design, the aim is to reproduce this precise organization of the blocks matrix using OpenCL code and the IntelFPGA OpenCL SDK, which generates a board organization and the full setups given an OpenCL code. In this article, we investigate the challenges for speeding up Big Data algorithms and provide a roadmap for further improvement. Ultimately, our aim is seamless integration of FPGAs to build a micro-server for Big Data processing. Many recent research has been conducted to strengthen the niche of FPGA in big data business [3], [4], [5]. However, the current available tool sets to build FPGAs are still complex. On the other hand, GPUs are leveraged well with the implementation of CUDA language. Besides, GPUs where instructions are executed in fixed instruction set are more flexible; they are also more adapted for floating point arithmetic.

The paper is organised as follow: Section II present the related work and our contribution, Section III presents our methodology applied for performance evaluation. In Section IV, we will give a description of the test benchmarks implemented in our work. Section V presents the hardware specification as well as the software tools and the adopted parallel programming technology. In Section VI we will present

the algorithms implementation as well as the performance evaluation. Section VII gives a holistic overview and conclude the work

## II. RELATED WORK AND CONTRIBUTION

Authors in [6] presented an FPGA-Accelerated Big Data implementation. The proposed system is based on the popular Apache Spark framework on the software side, and on an OpenCAPI-based POWER9 platform with Xilinx VU37P FPGA on the hardware side. Their system is able to generate a high-performing FPGA circuit from very high-level code descriptions in Spark.

The work in [7] present a case study of accelerating Apache Spark using re-configurable architecture. The authors proposed a framework to integrate FPGA accelerator into a Spark cluster. The Spark tasks are accelerated on the FPGA using Python. The performance results are evaluated with a case study of 2D FFT algorithm acceleration. The obtained results showed that FPGA based Spark implementation acquires 1.79x speedup than a conventional CPU implementation.

The author in [8] proposed the use of distributed databases and high-performance computing architecture in order to exploit multiple re-configurable computing and application specific processing. The proposed a 4-layer general architecture for smart agriculture, which is able to collect, store and process data from IoT nodes, integrate external data from other sources and allows efficient treatments of data coming from several sources with a cloud high-performance heterogeneous architecture.

A collaborated research team from George Mason University and University of California, guided by Netshatpour [3] has discovered a significant speedup with respect to K-means, KNN, SVM and Naive Bayes while implementing the mentioned machine learning algorithms in a Hadoop Platform with Intel Atom C2758 and Xeon E5 as master nodes and several Xilinx Zynq devices as slave nodes.

Besides, [4] have presented an FPGA-based hardware accelerator platform for big data matrix processing. A comparison of performance has been conducted between an Intel I7-4770 CPU (3.4GHz) and an FPGA of the model VC707 (125 MHz). Furthermore, a recent research published by the University of Science and Technology of China [5] has presented a software-defined operating system framework for FPGA based accelerator with the implementation on Xilinx Zynq FPGA.

On the other hand, ITRS Semiconductor roadmap foresees that hundreds of processors would be the base for the next generation embedded multicore designs. Recently, Microsoft in 2015 collaborated with Bing to investigate the use of FPGA. The project, also known as Project Catapult, has showed an improvement of nearly a factor of two of the operations per second in a critical component of Bing search engine [9].

With the demand for high speed network and computing, speed and parallel algorithms have become essential tools for development. Many of these operations were performed by a general purpose processor. But now days due to the availability of FPGAs, many researchers try to implement various algorithms on FPGAs more efficiently. FPGAs are often used

TABLE I. PERFORMANCE METRICS

| Metrics used for FPGA and GPU | Additional metrics for FPGA |
|-------------------------------|-----------------------------|
| Execution time                | % of DSP blocks             |
| Memory bandwidth              | % Logic Elements            |
| -                             | % of Memory Block           |

as hardware accelerators. Our work aims to implement big data benchmarks and to compare the performance of GPUs and FPGAs when it comes to big data processing. The main idea of our targets is to construct a scaled up platform of FPGA-based micro-server for big data processing. To date, among all the current research and works in the state of the art, none of them targets the application of FPGA in big data processing. To our knowledge, this is the first work to evaluate and optimize big data algorithms on a dedicated architecture by adopting the high level synthesis approach.

## III. PERFORMANCE EVALUATION METHODOLOGY

### A. Performance Metrics

The FPGA and the GPU we plan to use does not have the same I/O maximum speed, which will result in a comparison error if the slowest computing unit is limited by its I/O maximum bandwidth. Therefore, for each algorithm and each implementation, we plan to save the different metrics during the computing of the same series of test files on the two components. These metrics will then be compared by being put into charts and analysed.

For the GPU: We use Nvvp which is a built-in tool of Nsight to evaluate the performance of GPU. This tool allows us to measure the running time and the throughput of each function implemented, hence, allows us to know which parts of the code can be improved. For FPGA: Assuming that there is still no benchmark available regarding FPGA for the algorithms we decided to propose our own implementation. There is no implicit way of evaluating their performances precisely. Therefore, our study is based on a data-processing speed comparison between an FPGA and a GPU, the last being already used for massive parallel computing. The performances of our algorithms will also be verified by the IntelFPGA SDK for OpenCL, which includes an optimization report. The Table I gives some important metrics evaluated in our results.

### B. Multiobjective Optimization and Pareto Optimality

FPGA programming is about finding algorithms that optimize some aspects of the performance regarding different limited resources. Therefore, there is no unique solution to those problems, and the optimal computation depends on the factors that we want to optimize. No single FPGA implementation for all benchmarks can be an optimal  $\langle P, E, A \rangle$  with P for Performance, E for Energy and A for Chip Area [10]. Therefore, with  $N$  benchmarks each being potentially individually implemented Pareto like with numerous configurations the automatic optimization variations on concurrency E against A is needed.

TABLE II. EXECUTION OUTPUT AND DATASET STATISTICS

|                                 |             |
|---------------------------------|-------------|
| Average path distance           | 3.692507    |
| Network diameter                | 8           |
| Global efficiency               | 0.306578    |
| Clustering coefficient          | 0.632353    |
| Transitivity                    | % 0.000073  |
| Betweenness centrality          | 4051.734470 |
| Closeness centrality            | 0.261441    |
| Degree Distribution             | 0.023026    |
| Pearson correlation coefficient | -1.536665   |

### C. Graph Measures Results

The implemented big data algorithms were tested on the Stanford Large Network Dataset Collection provided by SNAP (Stanford Network Analysis Project) [11]. We have executed our algorithms on the dataset consisting of <circles> (and <friends lists>) from Facebook, which can be represented by an undirected graph of 4039 vertices and 88234 edges. Table II represent the execution outputs of the created graph. We have implemented the graph using C language and produce the algorithms to measure the dataset statistics: the Average path distance, Network diameter, Global efficiency, Clustering coefficient, Transitivity, Betweenness centrality, Closeness centrality, Degree distribution and Pearson correlation coefficient.

### D. Test-bed Setup

To evaluate the OpenCL FPGA implementation, we used a host computer, operating at 2.5 GHz under CentOS Linux 7.0 with a 32 GB RAM, with FPGA board mounted on the PCIe slot. We used the De5a-Net board embedding the IntelFPGA Arria 10, Fig. 1 shows the test-bed setup.

## IV. ALGORITHM DESCRIPTION

### A. K-means

K-means is a popularly-used algorithm for clustering. The aim of clustering is to divide the given set of data  $X$  composed of  $n$  points into partition  $\{C_i\}_{1 < i < k}$  such that points in each subset are similar to each other; otherwise, points from different groups are dissimilar [12]. The similarity is defined by a distance function; therefore, clustering task is able to be interpreted quantitatively as minimizing the cost function desired by user, which is normally formulated as below:

$$e_k(X, C) = \sum_{i=1}^n \min(D(x_i, c_j)) \quad (1)$$

where  $c_j$  is the center of subset  $C_j$

For K-means, we set the distance  $D(x, y)$  by the square of Euclidean distance  $\|x - y\|^2$  (this is not a metric, because it does not have triangular inequality property). Hence, K-means cost function is defined by:

$$e_k(X, C) = \sum_{i=1}^n \min_{1 < k < j} \|x_i - c_j\|^2 \quad (2)$$

Given a set of data points, a clustering algorithm aims to the similarity which is defined using distance measure. In our

work, the Euclidean distance is utilized. Basically, we start by choosing  $K$  points called centroids randomly among the given points then form  $K$  clusters, each of which contains a centroid and the points that accept this centroid as the nearest one. We gradually update these centroids by calculating the center of mass in each group. This algorithm terminates when the number of iterations exceeds a chosen number or the change after each iteration is less than a chosen threshold

1) *Lloyds heuristic algorithm for Kmeans*: The Lloyds heuristic algorithm [13] for clustering high-dimensional data is usually described by 4 steps. Firstly, stop condition is defined as following: algorithm terminates after exceeding a number of loops or whenever the difference of  $e_k(X; C)$  between two consecutive loops is less than a real positive threshold  $r$  given.

- Step 1: Initialize  $k$  temporary centroids. Start loop of  $N$  iterations:
- Step 2: For each  $x$  in given data set, search for the nearest centroid  $c$  from  $x$  and assign  $x$  to this cluster.
- Step 3: For each cluster  $C_i$ , calculate the new centroid of  $C_i$  by following formula :

$$c_i = \frac{1}{C_i} \sum_{x \in C_i} x \quad (3)$$

- Step 4: Calculate the new value of  $e_k(X; C)$ , then the difference  $\Delta = e_k(X; C)^{new} - e_k(X; C)^{old}$ . If  $\Delta$  is smaller than  $r$ , return the contemporary assignment and end the loop.

### 2) Initialization Method:

a) *Method 1 (Forgy [14])*: : Choose arbitrarily  $k$  points from data set, this method gives us no guarantee about how close the cost function will be to the global minimum. Therefore, to increase the chance to get well-accepted result, we repeat this initialization  $l$  times and pick out which gives the best output.

b) *Method 2 (K-means++)*: : This method guarantees that:

$$E[e_k] \in 8(2 + \ln k)e$$

where  $e$  is the global minimum, hence allows us to control the performance of heuristic algorithm.

---

### Algorithm 1 K-means++

---

Choose  $c_1$  uniformly from data set:  $C \leftarrow \{c_1\}$

**for**  $i = 2$  **to**  $k$  **do**

    Choose  $c_i = x \in X$  with the probability

$$p(x) = \frac{D^2(x, C \cup \{x\})}{\sum_{y \in X} D^2(y, C \cup \{x\})}$$

$C \leftarrow C \cup \{c_i\}$

**end**

---

### B. Sorting Network

One of the commonly used operations in high speed data processing is data sorting. A sorting network consists of two types of items: comparators and wires. The wires are



Fig. 1. Test-bed Architecture (DE5aNet board).

thought of as running from left to right, carrying values (one per wire) that traverse the network all at the same time. Each comparator connects two wires. When a pair of values, traveling through a pair of wires, encounter a comparator, the comparator swaps the values if and only if the top wire value is greater than the bottom wire value. Sorting networks differ from general comparison sorts in that they are not capable of handling arbitrarily large inputs, and in that their sequence of comparisons is set in advance, regardless of the outcome of previous comparisons. This independence of comparison sequences is useful for parallel execution and for implementation in hardware. Some well-known methods to construct a sorting network can be listed such as Batcher odd-even merge sort [15], bitonic sort [16], Shell sort [17] and the Pairwise sorting network [18], whose depth efficiency is  $O(\log^2(n))$ . Basically, the sequential sorting algorithm requires at least comparisons. To boost the performance on treating massive volume of data, some specified algorithms are chosen to be parallelized depending on their characteristics (data dependency, device architecture, methods of communication, network topology, etc.). The most commonly used sorting algorithm is Bubble sorting. For efficient and reduced operations implementation of sorting, [15] proposed a technique of sorting using sorting networks.

1) *Bitonic Sort Algorithm:* The bitonic sort [19] is a divide and-conquer comparison sort usually implemented with recursion. Keys are first ordered into bitonic sequences and are then sorted using a bitonic merger. The number of comparators required can be reduced by a factor of  $\log^2(N)$  by combining the perfect shuffle with bitonic sorting. This sort fits the SIMD (single instruction multiple data) model because it is readily implemented in hardware using a parallel sorting network. Given sufficient hardware, this sort is capable of achieving  $O(\log^2(N))$  performance. A sequential, recursive version of the bitonic sort with running time  $O(N \log^2(N))$  is used on the microprocessor. The FPGA implementation uses a visualized, parallel sorting network. Both implementations sort in-place and require the input key quantity to be a power of 2; however, on the FPGA, we were able to use the same SIMD controller to schedule keys for an eight input sorting network

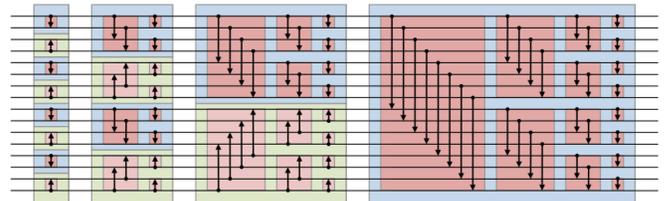


Fig. 2. Bitonic Sorting Network Illustration .

and a four input sorting network simultaneously. This allowed the use of all six memory banks. Bitonic sortings goal is to sort a bitonic sequence (a sequence  $(a_n)$  is called bitonic if and only if there exist an unique index  $i$  such that for all integer  $m$ , if  $a_m; a_{m+1}; \dots; a_{m+i}$  are monotonically increasing (or decreasing), then  $a_{m+i+1}; \dots; a_{m+2i}$  must be monotonically decreasing (or increasing)), which is easy to be parallelized and applied on hardware or software. For an arbitrary sequence, we can sort the first half of sequence in increasing order and the other in decreasing order, this transforms the sequence into a bitonic sequence in order to apply bitonic sort. For each  $i$  from 1 to  $n = 2$ , match two elements  $x_i$  and  $x_{i+n/2}$  into a pair and swap them to form a pair in order (min; max). Consequently, we obtain two bitonic sequences whose lengths are a half of the given one (bitonic property is invariant by step 1) and also every element in the first half is smaller than every the second half. Apply recursively this procedure to each of two sub sequences until the length of sequence is less than 2.

### C. Correlation Algorithms

Correlation algorithms such as Pearson [20] and Spearman [21] measure the dependence between variables. Popular applications may include calculating the relationship between age and number of hours spent watching TV, or the relationship between product sales and temperature. Assuming that comparisons and simple operations are both done in time  $O(1)$ , Pearson and Spearman correlations are done respectively in time  $O(n)$  and  $O(n \log(n))$ . The Pearson correlation coefficient is given by the formula (4), assuming that  $a$  and  $b$  are

two zero-mean real valued random variables.

$$\rho(a, b) = \frac{E(a, b)}{\sigma_a, \sigma_b} \quad (4)$$

where  $E(a, b)$  is the cross-correlation between  $a$  and  $b$ , and  $\sigma_a^2 = E(a^2)$  and  $\sigma_b^2 = E(b^2)$  are the variance of  $a$  and  $b$  respectively. It is more convenient to work with the squared Pearson correlation coefficient given by (5)

$$\rho^2(a, b) = \frac{E^2(a, b)}{\sigma_a^2, \sigma_b^2} \quad (5)$$

The squared Pearson correlation coefficient give an insight about the strength of the linear relationship between two random variables. When  $\rho^2(a, b) = 0$ , then two random variables  $a$  and  $b$  are uncorrelated. When the value of  $\rho^2(a, b)$  is near to 1, then  $a$  and  $b$  are said to be correlated. The squared Pearson correlation coefficient detects only linear dependencies between the two variables  $a$  and  $b$ . Indeed, If  $a$  and  $b$  are independent, then  $\rho^2(a, b) = 0$ , but the converse is not true. The Pearson correlation coefficient  $\rho_p$  is defined according to equation 6:

$$\rho_p = \frac{\sum_{i=1}^N a_i, b_i}{\sqrt{\sum_{i=1}^N a_i^2 \sum_{i=1}^N b_i^2}} \quad (6)$$

The Spearman correlation coefficient  $\rho_s$  is calculated in the same manner as  $\rho_p$ , except that  $\rho_s$  is calculated after both  $a$  and  $b$  have been rank transformed to values between 1 and  $N$  (Equation 7). When calculating  $\rho_s$ , a fractional ranking is used, which means that the mean rank is assigned in case of ties. For example, suppose that the two smallest numbers of  $a$  are equal, then they will be both ranked as  $1.5 (\frac{1+2}{2})$ . A mean centering is first performed (by subtracting  $N/2 + 1/2$  from each of the two ranked vectors).

$$\rho_s = \frac{\sum_{i=1}^N a_{i,r}, b_{i,r}}{\sqrt{\sum_{i=1}^N a_{i,r}^2 \sum_{i=1}^N b_{i,r}^2}} \quad (7)$$

## V. HARDWARE SPECIFICATION

### A. Field Programmable Gate Arrays

1) *Arria 10 Architecture:* As a dedicated architecture, We used the Arria 10 FPGA as a target platform for our algorithm implementation Fig. 3. Arria 10 is one of the latest chip produced by IntelFPGA delivering the highest performance at 20 nm. Arria 10 FPGA is a low power embedded architecture up to 40% lower power than previous FPGAs generation. It allows up to 1500 GB/s floating-point operation with DSP blocks. The system clock is 100 MHz. The chip also includes a Dual-Core ARM operating at 1.5 GHz. Table III shows the available resources in terms of logic elements, DSP and memory blocks of the Arria 10 FPGA.

2) *High Level Synthesis:* The Arria 10 FPGA programming is done using OpenCL based High Level synthesis [22]. This is to achieve an efficient and fast parallel implementation of the algorithm on FPGA. OpenCL (Open Computing Language) is the first open, royalty-free, unified programming model for accelerating algorithms on heterogeneous systems. Based on

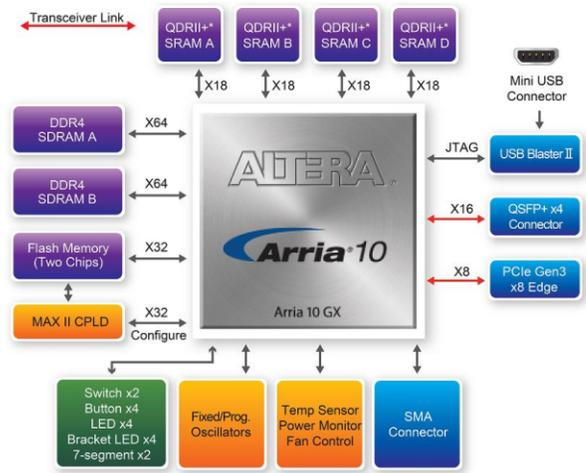


Fig. 3. Arria 10 Architecture .

TABLE III. RESOURCES OF ARRIA 10

| Resource                | Arria 10 device<br>10AX115N2F45E1SG         |
|-------------------------|---------------------------------------------|
| Logic Elements (LE) (K) | 1,150K                                      |
| ALM                     | 251,680                                     |
| Register                | 1,006,720                                   |
| Memory                  | 32MB QDRII+ SRAM<br>16GB DDR4 SO-DIMM SDRAM |
| DSP Blocks              | 1,518                                       |
| 18 x 19 Multiplier      | 3,036                                       |
| 17.4 Gbps Transceiver   | 48                                          |
| PCIe Hard IP Block      | 4                                           |
| Embedded memory         | 67-Mbits                                    |

C (C99), it supports four kinds of processing units: CPU, GPU, FPGA and DSP (digital signal processors). The real asset of this language is to use different units at the same time, processing them in parallel, and using each one of them for what it is the best. However, because of the total differences in processing algorithms between the different sorts of units, an OpenCL code has to be optimized for each device. The IntelFPGA SDK for OpenCL allows avoiding the traditional hardware FPGA development, which is too complicated for the use when it comes to high performance computing, in order to achieve a much faster and higher level software development flow. It includes multiple optimizations and can produce deep reports of the compilation and the code optimization. The IntelFPGA SDK for OpenCL requires the Quartus Prime (Pro version for Arria 10 board), also known as Quartus II, to function optimally. Fig. 4 shows the compilation process of OpenCL code for FPGA.

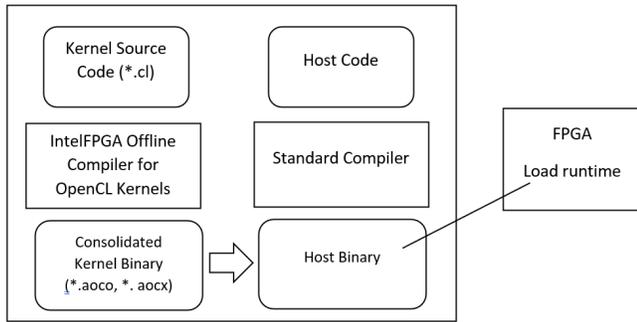


Fig. 4. The Flowchart of Compilation Process of the FPGA .

3) *Quartus TCL Scripting*: The Quartus II development software provides the scripting environment, particularly Tcl (tool command language) scripting. We use scripting support, to achieve custom analysis, automation and reproducibility. Custom analysis allows to build test procedures into the script and change design processing based on the test results. Scripts can automate design flows to perform on the computer and easily archive and restore projects. Reproducibility ensures that scripts use the same project setup and assignments for every compilation.

### B. General Purpose Graphical Processing Unit GPGPU

1) *GPU Architecture*: In our work, we used the Nvidia GPU Quadro K2200 (Table IV). This GPU uses the first generation of Maxwell architecture released by Nvidia in February 2014 (the newest and second generation was released in September 2014). Maxwell introduces an all-new design for the Streaming Multiprocessor (SM) called SMM that dramatically improves energy efficiency compared to its predecessor Kepler. SMM uses a quadrant-based design with four 32-core processing blocks each with a dedicated warp scheduler capable of dispatching two instructions per clock. Each SMM provides eight texture units, one polymorph engine (geometry processing for graphics), and dedicated register file and shared memory. Maxwell improves on Kepler by separating shared memory from L1 cache, providing a dedicated 64KB shared memory in each SMM (for Quadro K2200). It provides native shared memory atomic operations for 32-bit integers and native shared memory 32-bit and 64-bit compare-and-swap (CAS), which can be used to implement other atomic functions.

2) *Programming the GPGPU*: CUDA is a parallel computing platform and application programming interface (API) model created by Nvidia. It allows software developers to use a CUDA-enabled graphics processing unit (GPU) for general purpose processing an approach known as GPGPU. The CUDA platform is a software layer that gives direct access to the GPU virtual instruction set and parallel computational elements. The CUDA platform is designed to work with programming languages such as C, C++ and Fortran. Quadro K2200 uses the version CUDA 5.0. We are using Nsight Eclipse Edition (CUDA SDK 6.0) for algorithm implementation. This is a full-featured IDE powered by the Eclipse platform that provides an all-in-one integrated environment to edit, build, debug and profile CUDA-C applications.

## VI. ALGORITHM IMPLEMENTATION AND PERFORMANCE EVALUATION

### A. FPGA Design and OpenCL Optimization

The conception of the design is a crucial step in the development process. We dispose of four types of blocks (Logic blocks, Memory blocks, Logic Register and Digital Signal Processing Blocks). Each one is able to perform a particular list of actions. Given these four sorts of blocks and their amount on the board, the aim is to associate to each one of them simple and statics instructions to do and to link them in a network. For instance, if we want to compute the sum of four 64-bytes integers  $i_1$ ,  $i_2$ ,  $i_3$  and  $i_4$ , we chose to dedicate:

- One Logic Blocks programmed to do the summation of  $i_1$  and  $i_2$ , then send the data directly to the third block.
- One Logic Blocks programmed to do the summation of  $i_3$  and  $i_4$ , then send the data directly to the third block.
- One Logic block programmed to do the summation of the two previous results

On this example, we see an important characteristic of FPGA: the temporary variable of summation doesnt have to be stored, so there is no need for write/read operations, reducing the total amount of clock-cycles. If we want to be able to compute a new set of data every clock-cycle, the path taken by the data has to be the same regardless of the data themselves, even though we increase the number of comparison or assignment.

OpenCL is a language developed in order to support multi-platform computing. Considering the deep differences between computing units (CPU,GPU), the same OpenCL code is implemented highly differently on different platforms on a hardware level. Therefore, even though an OpenCL code for GPU will work on other platforms, it will only be optimized for GPU, and will certainly be highly inefficient if run on other computing units. It is the same for FPGAs: multiple valid implementations are inefficient, and there are multiple ways of coding that have to be promoted or avoided.

The first coding optimization specific to IntelFPGA OpenCL is the command `#pragma unroll` that has to be put in the OpenCL kernel file. Used before a loop, this command is read by the compiler, that process what is called the unrolling of the loop. On a hardware level, each iteration of a nonunrolled loop is by default done by the same area of the computing unit. The process of unrolling by a factor N consists in replicating the hardware  $(N - 1)$  time in order to be able to compute N iteration of the loop at the same time. If it seems efficient to unroll the loops, it is important to notice that the factor of the unroll has to be defined during the synthesis step, and cant be changed during the computation. Moreover, the hardware resources being limited, before unrolling it is important to be sure that the concerned loop results in a bottleneck of the overall computing process. It is useless to unroll a loop more than the number of iterations, and difficult to unroll it if the number of iterations is not easy to determine. It is important to underline the differences between the two kinds of parallelization that we have presented: the loop unrolling is a hardware-parallelization, whereas the ability

TABLE IV. SPECIFICATIONS FOR THE QUADRO K2200

|                 | Processing Power<br>(GFLOPS)                           | Memory Clock<br>(MHz) | Memory    |                   |          |                     |
|-----------------|--------------------------------------------------------|-----------------------|-----------|-------------------|----------|---------------------|
|                 |                                                        |                       | Size (MB) | Bandwidth<br>GB/s | Bus type | Bus width<br>(bits) |
| Quadro<br>K2200 | 1280 (Single<br>precision)<br>40 (double<br>precision) | 1250 (5000)           | 4096      | 80                | GDDR5    | 128                 |

of computing a new set of data every clock-cycle is a time-parallelization. Their combination can theoretically lead to drastic computing acceleration.

In order to simplify the unrolling process, it is important to avoid nested loops as much as possible. A nested loop is a loop called in another loop. The more loops inside a loop there are, the harder it is to unroll them. Therefore, an OpenCL code should promote one, maximum two loop levels with explicit amount of iterations. For these reasons, the OpenCL kernel for our algorithms are implemented with only two levels of loops having an explicit amount of iterations.

Another optimization possible to perform is the balance of reducing a set of data using an associative operation. For instance, if you want to add a set of  $N$  integers, the usual loop unrolled will use  $N-1$  Logic Blocks. However, by doing a tree summation, the number of blocks can be reduced to  $N/2$ . This optimization can be managed by the compiler, even for more complex operations. Moreover, if the overall computation tree implies non-associative operations, the compiler can identify parts of the computation tree that can be balanced and balance them. Because of the difficulty in identifying such parts, the compiler proposes this optimization as an option. It is to the programmer to understand if this auto-balancing function is relevant, or to balance manually by modifying the code when the compiler cant extract the balancing. Many other constrains and way of coding are to consider when producing OpenCL code for FPGA, like the impossibility to use pointer to pointer parameter in the kernel functions, or the simplicity of indexes when arrays are called. In order to show the efficiency of OpenCL kernel optimization, we have implemented two version of Summary Statistics algorithm: Optimized (Algorithm 2 ) and unoptimized (Algorithm 3) kernels. This algorithm calculate the column-wise min, max, mean, variance, count, and number of non-zeros in a given dataset. (assuming each simple is done in  $O(1)$ , all these statistics share the same complexity of  $O(n)$ ).

Table V shows the estimated resource usage before kernel optimization. The problem reported by the optimization report is that too many kernels attempted to access the same variable at the same time (hereby is the variable sum), but there is only limited amount of access possible on the same variable each clockcycle. Therefore, the blocks that try to access have to wait, blocking the overall process and retarding it by  $N$  clock-cycle. An efficient way to avoid this problem is to use a shift-register with the size  $N$ , the maximum encountered late. The amount of clock-cycle is revealed by the optimizer.

Table VI shows the estimated resource usage after kernel optimization. The kernel optimization has improved the computation speed. The processing time is divide by a factor  $N$ , which is the number of clock-cycles that were lost because

TABLE V. ESTIMATED RESOURCE USAGE SUMMARY (UNOPTIMIZED SUMMARY STATISTICS)

| Resource                  | usage |
|---------------------------|-------|
| Logic utilization         | 16%   |
| Dedicated Logic registers | 8%    |
| Memory blocks             | 28%   |
| DSP blocks                | 4%    |

of blocking access. The IntelFPGA optimization report ensure to identify the most important bottlenecks and processes that slow down the computing and increase the number of clock-cycles taken by an overall computation. These optimizations strategy are adopted for the other algorithm in order to achieve an efficient parallel implementation with less resources usages.

**Algorithm 3** Unoptimized OpenCL kernel

```
__kernel void summarystat (__global float *A,
unsigned int size, __global float*rep, __global
int*non_zero)
```

```
min = A[0]; max = A[0]; sum = 0.0f; sqsum = 0.0f;
nonzero_count = 0;
```

```
#pragma unroll
for (i = 1; i < size; i++) do
 x = A[i];
 sum += x;
 sqsum += x * x;
 if x ≠ 0 then
 | nonzero_count ++
 end
 if x < min then
 | min = x
 end
 if x > max then
 | max = min
 end
end
```

```
rep[0] =max;
rep[1] =min;
rep[2] =sum/size;
rep[3] =sqsum/size - (sum/size)2;
*non_zero = nonzero_count;
```

**B. OpenCL Implementations of Bitonic Sort Algorithm**

If we consider a regular optimal sorting algorithms like the Quick sort, which has a  $O(n \log(n))$  complexity in term of

**Algorithm 2** Optimized OpenCL Kernel

```
__kernel void summarystat_optimized(__global
float *restrict A, unsigned int size, __global
float*restrict rep, __global int*restrict non_zero)
min_temp = A[0]; max_temp = A[0]; min_rep = min_temp;
max_rep = max_temp; sum = 0.0f; sqsum = 0.0f;
nonzero_count = 0;
```

```
float shift_reg_x_s[N + 1]; float shift_reg_sqx[N + 1];
float shift_reg_min[N + 1]; float shift_reg_max[N + 1];
int shift_reg_non_zero[N + 1];
for (i = 0; i < N + 1; i++) do
 shift_reg_x_s[i]=0;
 shift_reg_sqx[i]=0;
 shift_reg_min[i]=min_temp;
 shift_reg_max[i]=max_temp;
 shift_reg_non_zero[i]=0;
```

```
end
for (i = 0; i < size; i++) do
 x= A[i]; shift_reg_x_s[N]= shift_reg_x_s[0] + x;
 shift_reg_sqx[N] = shift_reg_sqx[0] + x * x; if x <
 shift_reg_min[0] then
 | shift_reg_min[N]=x
 end
 else
 | shift_reg_min[N] = shift_reg_min[0]
 end
 if x < shift_reg_max[0] then
 | shift_reg_max[N]=shift_reg_max[0]
 end
 else
 | shift_reg_max[N] = x
 end
 if x ≠ 0 then
 | shift_reg_non_zero[N] = shift_reg_non_zero[0] + 1
 end
 #pragma unroll for (j = 0; j < size; j++) do
 | shift_reg_x_s[j]= shift_reg_x_s[j + 1]
 | shift_reg_sqx[j]= shift_reg_sqx[j + 1]
 | shift_reg_min[j]= shift_reg_min[j + 1]
 | shift_reg_max[j]= shift_reg_max[j + 1]
 | shift_reg_non_zero[j]= shift_reg_non_zero[j + 1]
 end
end
#pragma unroll for (j = 0; j < size; j++) do
 sum +=shift_reg_x_s[j]; sqsum += shift_reg_sqx[j];
 nonzero_count += shift_reg_non_zero[j]; if min_rep >
 shift_reg_min[j] then
 | min_rep = shift_reg_min[j]
 end
 if max_rep > shift_reg_max[j] then
 | max_rep = shift_reg_max[j]
 end
end
end
rep[0] = max_rep; rep[1] = min_rep; rep[2]= sum/size;
rep[3]=sqsum/size - (sum/size)2; *non_zero = nonzero_count;
```

TABLE VI. ESTIMATED RESOURCE USAGE SUMMARY (OPTIMIZED SUMMARY STATISTICS).

| Resource                  | usage |
|---------------------------|-------|
| Logic utilization         | 8%    |
| Dedicated Logic registers | 4%    |
| Memory blocks             | 10%   |
| DSP blocks                | 1%    |

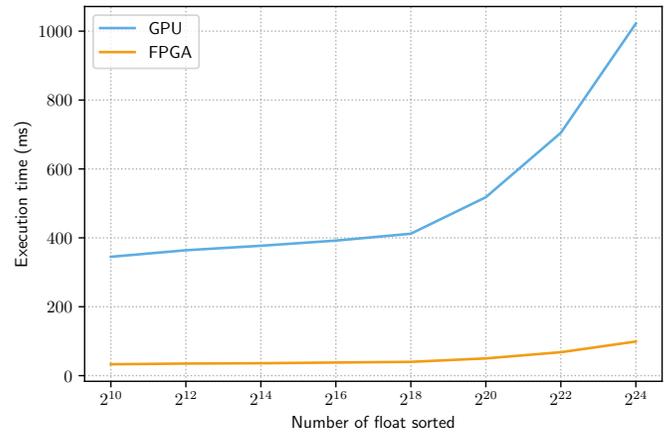


Fig. 5. Execution time of the Sort Algorithm for Many Input Sizes.

comparisons, it seems impossible to find a fixed organization of blocks that applies the algorithm to any input data. Whereas sorting networks, like Bitonic Sort present a complexity of  $O(n \log^2(n))$ , seems highly optimized for FPGA, because they use a fixed data path. For the Bitonic sort algorithms, we configure the logic blocks to have two input and two output: the first one returning the max of the two input and the second one returning the minimum of these two. With these sorting networks, each new list to sort is computed every  $p$  clock cycles,  $p$  being the number of clock-cycle taken by logic block to return the max and min of the two inputs. In order to achieve a parallel implementation of Bitonic sort we divide the dataset into multiple threads (one thread occupies at least one data). For each step of bitonic sort as we can notice in Fig. 2, all of comparing operations are executed simultaneously on available threads. Fig. 5 shows the performance evaluation of sort implementation on the GPU and FPGA. We run the GPU and FPGA implementation to sort a set of different float ranging from  $2^{10}$  to  $2^{24}$ . The obtained results shows the parallel computing power of the FPGA. For even large number of sorted float ( $2^{24}$ ) the FPGA implementation is always efficient compared to the GPU implementation which need more then 1 second to process ( $2^{24}$ ) floats.

**C. OpenCL Implementations of K-means Algorithm**

K-means has the difference of data independence from Bitonic Sort. In fact, to find the nearest centroid from one point, we can assign each points from data set onto available threads and perform the calculation and comparison; on the other hand, new centroids calculation needs data from the membership matrix, this step can be parallelized by assigning

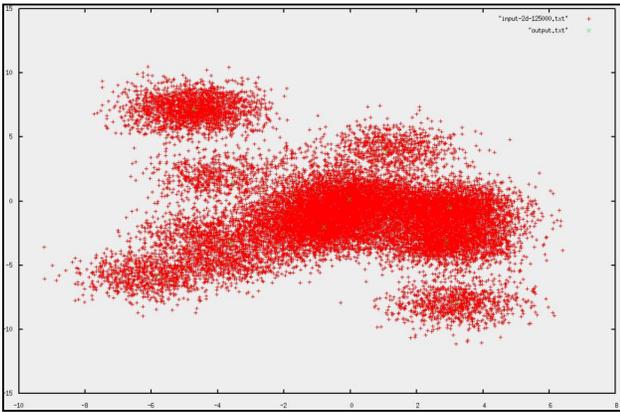


Fig. 6. K-mean GPU Result of Clustering 125000 2D-points into 10 Clusters.

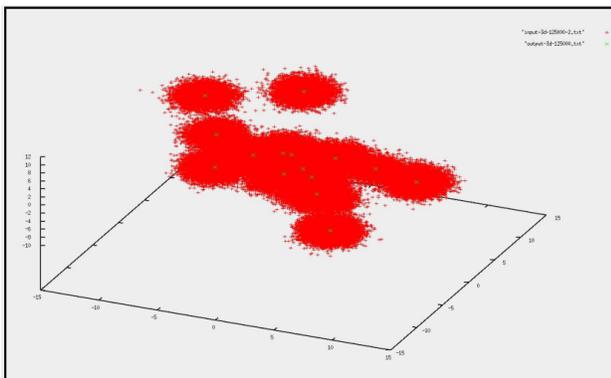


Fig. 7. K-mean FPGA Result of Clustering 125000 3D-points into 15 Clusters .

each cluster to a thread and carrying out the calculation separately for each of them. However in this implementation, we let the host device take the task sequentially. For K-mean implementation, the test data is generated by the Spark benchmark. Firstly, we tested the algorithm on small dimensional data (2D and 3D) in order to test the output. Then the algorithm is tested on 10 and 100 dimension points sets. Finally, we tested on high dimensional data (500D). The size of data ranged from 12500 to 125000 points for small dimensional sets and up to 31250 points on high dimensional sets (up to 300MB). Figures Fig. 6 and Fig. 7 shows respectively the GPU and FPGA results of the k-mean implementation. Each red point is the visualization of a point in the given data set and each green point is a centroid found by the heuristics. The obtained results confirm the functional validation about clustering. Indeed, the cost function (sum of square of the distance from each point to the centroid of the cluster containing it) is shown to decrease after each iteration.

Table VII shows the performance evaluation of k-mean implementation on the GPU and FPGA. The running time to cluster 125000 3D points is 380 ms for the GPU implementation and 33 ms for the FPGA implementation. For 31250 500D points is 12 seconds on the GPU and 158 ms on the FPGA. The occupancy achieved is 98.8 percent in the first measurement and 49.5 percent (over the theoretical 50 percent) in the second. This shows that the occupancy is well controlled. The GPU

TABLE VII. GPU AND FPGA K-MEAN EXECUTION TIME

| Execution Times |              |              |
|-----------------|--------------|--------------|
| Platform        | 125 000 (3D) | 31250 (500D) |
| GPU             | 380 (ms)     | 12 (s)       |
| FPGA            | 33 (ms)      | 158 (ms)     |

implementation is far from real time performances. Indeed, the choice of block and grid size strongly affects the efficiency. For example, if the block size is reduced from (256 x 1 x 1) to (64 x 1 x 1), the time to find the centroids for the 125000 3D point set goes down from 380 ms to 250 ms. This can be explained by the use of synchronization in each block. By contrast, if the block size is increased from (61 x 1 x 1) to (1024 x 1 x 1) the running time on the data set of 31250 500D points goes down significantly from 12 seconds to 1.9 seconds. This can be explained by the large size of shared memory in each block. In the GPU implementation, the low DRAM utilization may come from the non-coalesced access to memory. Reducing this can also result in the better performance. The FPGA implementation is shown to outperform the GPU one. Only 158 ms is needed to cluster a very high dimensional data (500D).

#### D. OpenCL Implementations of Correlation Algorithms

The Pearson Correlation Algorithm was implemented using two reductions to find the mean of both input vectors, followed by the computation of the covariance of both entries and each one of the standard deviations. The final result is given by  $cov(X; Y) / (\sigma_X * \sigma_Y)$ . To calculate the Spearman coefficient we need to calculate the Pearson coefficient of the ranks. For the computation of the ranks, we first sort both samples by the values of the first one and calculate the ranks for the first sample by using a simple kernel that for each position in the sample that has an element different from the next one, goes back and counts all the occurrences of that element and then finally fills all the position with that same value with the mean of the ranks. This part of the code is not very parallelizable and can run in  $O(n)$  if all the elements in the initial sample are the same, opposed to  $O(1)$  with all elements different from one another [23]. But since in normal samples with float values this is unlikely to happen the approach works well. Then we sort again the first, the second and the ranks of the first by the values of the second sample, calculate now the ranks of the second sample using the same method and now that we got both ranks array, we use the Pearson Correlation Coefficient algorithm in this data to find the Spearman Correlation Coefficient. For measuring the execution time of both correlation algorithms, we first generated 5 input files for different sizes of samples, then we ran each one of those input files and took the mean of the execution times for each one of the 5 input files. Fig. 8, Fig. 9 shows the performance evaluation respectively of the Pearson and Spearman Correlation algorithm implementation on the GPU and FPGA. We test both implementation using a number of elements in sample ranging from  $2^{10}$  to  $2^{24}$ . For a high number of elements  $2^{24}$  the GPU process the pearson correlation algorithm in 1 seconds and the Spearman algorithm in 5.5 seconds. For the same number of elements, the FPGA implementation process the pearson and spearman correlation in 90 ms and 400 ms respectively. The processing time of the correlation algorithms on the FPGA implementation

has been decreased by a factor x12 compared to the GPU implementation.

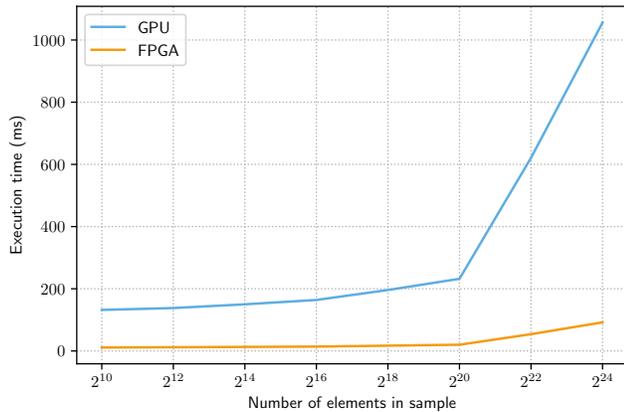


Fig. 8. Execution Time of the Pearson Algorithm for Many Input Sizes .

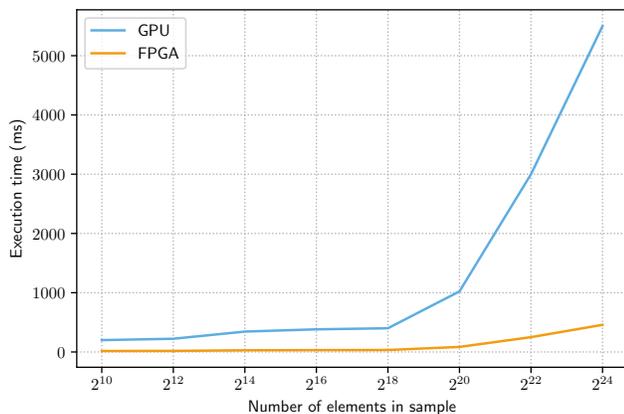


Fig. 9. Execution Time of the Spearman Algorithm for Many Input Sizes .

### E. Synthetic Results

Our work adopted the high level synthesis for FPGA implementation using OpenCL. Despite of the advantage of high level programming, its use is still limited. The Intel offline compiler takes a lot of time in order to generate the hardware configuration files (aocx) (duration of compilation hours for more complex function and if more optimization are requested from the compiler). However, we can achieve higher acceleration using OpenCL, which provides better memory management. We can freely access the local, global and constant memory in the OpenCL kernel. This allows us to better manage the data transmission and data structure. In addition, the FPGA is considered as the generation of integrated circuit claimed higher performance and reliability, and the emerging high level software tools make it easily accessible to the community. The encouraging results we obtained on the FPGA in term of acceleration performance, demonstrates that a dedicated architecture can be used to prototype a micro-server for big data that operates under real-time constraints.

As a future work, we intend to achieve a full embedded implementation for Big Data algorithms on FPGA using the integrated ARM processor of the Arria 10 SoC.

## VII. CONCLUSION

In this work, we have implemented and optimized three algorithms: Bitonic Sorting network, K-means, Spearman and Pearson correlation. The purpose behind this implementation is to prototype a micro-server for processing big data algorithms on both GPU and FPGA and compare their performance. We have implemented and quantitatively evaluated the execution times of some of the most important algorithms for big data. We present performance results on a heterogeneous architectures: high-end CPU-GPU and a dedicated CPU-FPGA architecture. The choice of using dedicated architecture was made principally because the big data algorithms can be massively parallelized. This property is exploited by using a dedicated FPGA-based architecture as a target platform for an efficient embedded micro-server. The performance of the optimized algorithms on the FPGA show a promising prospect of utilizing them in solving real-world problems.

## REFERENCES

- [1] A. K. Tiwari, H. Chaudhary, and S. Yadav, "A review on big data and its security," in *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*. IEEE, 2015, pp. 1–5.
- [2] D. Laney *et al.*, "3d data management: Controlling data volume, velocity and variety," *META group research note*, vol. 6, no. 70, p. 1, 2001.
- [3] K. Neshatpour, M. Malik, M. A. Ghodrati, and H. Homayoun, "Accelerating big data analytics using fpgas," in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2015, pp. 164–164.
- [4] C.-C. Chung, C.-K. Liu, and D.-H. Lee, "Fpga-based accelerator platform for big data matrix processing," in *2015 IEEE International Conference on Electron Devices and Solid-State Circuits (EDSSC)*. IEEE, 2015, pp. 221–224.
- [5] C. Wang, X. Li, and X. Zhou, "Soda: Software defined fpga based accelerators for big data," in *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2015, pp. 884–887.
- [6] J. Hoozemans, J. Peltenburg, F. Nonnemacher, A. Hadnagy, Z. Al-Ars, and H. P. Hofstee, "Fpga acceleration for big data analytics: Challenges and opportunities," *IEEE Circuits and Systems Magazine*, vol. 21, no. 2, pp. 30–47, 2021.
- [7] J. Hou, Y. Zhu, L. Kong, Z. Wang, S. Du, S. Song, and T. Huang, "A case study of accelerating apache spark with fpga," 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)*. IEEE, 2018, pp. 855–860.
- [8] O. Debauche, S. A. Mahmoudi, S. Mahmoudi, and P. Manneback, "Cloud platform using big data and hpc technologies for distributed and parallels treatments," *Procedia Computer Science*, vol. 141, pp. 112–118, 2018.
- [9] A. Putnam, A. M. Caulfield, E. S. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmailzadeh, J. Fowers, G. P. Gopal, J. Gray *et al.*, "A reconfigurable fabric for accelerating large-scale datacenter services," in *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*. IEEE, 2014, pp. 13–24.
- [10] Y. Censor, "Pareto optimality in multiobjective problems," *Applied Mathematics and Optimization*, vol. 4, no. 1, pp. 41–59, 1977.
- [11] J. Leskovec and A. Krevl, "Snap datasets: Stanford large network dataset collection," 2014.

- [12] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 503–527, 2007.
- [13] R. Ostrovsky, Y. Rabani, L. J. Schulman, and C. Swamy, "The effectiveness of lloyd-type methods for the k-means problem," *Journal of the ACM (JACM)*, vol. 59, no. 6, pp. 1–22, 2013.
- [14] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering," in *ICML*, vol. 98. Citeseer, 1998, pp. 91–99.
- [15] K. E. Batcher, "Sorting networks and their applications," in *Proceedings of the April 30–May 2, 1968, spring joint computer conference*, 1968, pp. 307–314.
- [16] K. J. Liszka and K. E. Batcher, "A generalized bitonic sorting network," in *1993 International Conference on Parallel Processing-ICPP'93*, vol. 1. IEEE, 1993, pp. 105–108.
- [17] M. T. Goodrich, "Randomized shellsort: A simple oblivious sorting algorithm," in *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 2010, pp. 1262–1277.
- [18] I. Parberry, "The pairwise sorting network," *Parallel Processing Letters*, vol. 2, no. 02n03, pp. 205–211, 1992.
- [19] D. E. Knuth, "Sorting and searching," 1973.
- [20] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.
- [21] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Encyclopedia of statistical sciences*, vol. 12, 2004.
- [22] J. Cong, B. Liu, S. Neuendorffer, J. Noguera, K. Vissers, and Z. Zhang, "High-level synthesis for fpgas: From prototyping to deployment," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 30, no. 4, pp. 473–491, April 2011.
- [23] S. Kim, M. Ouyang, and X. Zhang, "Compute spearman correlation coefficient with matlab/cuda," in *2012 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2012, pp. 000 055–000 060.

# Assessing and Proposing Countermeasures for Cyber-Security Attacks

Ali Al-Zahrani

College of Computer Sciences and Information Technology  
King Faisal University Al-Ahsa 31982  
Saudi Arabia

**Abstract**—Cyber-attacks on IT domain infrastructure directly affect the security of businesses' operational processes, potentially leading to system failure. Some industries have a high risk than others due to the sensitivity of their data, including the transportation industry, which has recently moved from traditional data management to digitalization. This study aims to identify the main cyber threats in the transportation sector by analyzing related works and highlighting the main countermeasures used to respond to such threats as well as enhance overall cybersecurity. This paper presents a comprehensive cybersecurity risk assessment for the transportation companies, identifying the most common attacks and proposing methods to minimize risk as much as possible. A risk assessment analysis was prepared by industry experts that included previous cyberattack scenarios. The results of our paper identified the most critical attacks on the transportation company's booking system and recommended suitable countermeasures to minimize the risk of those attacks.

**Keywords**—Cyber-attacks; cyber-security; risk assessment; countermeasures

## I. INTRODUCTION

### A. Background

Cyber-security attacks are considered one of the hot topics in the field of information security and can result in huge losses to organizations if not carefully handled. Cybersecurity attacks usually result from several factors related to threats, human errors or insufficient knowledge [1]. Cybersecurity relates to technologies, processes, practices, and information assets, aiming to protect against any damage or unauthorized access caused by cyberattacks [2]. Cyberattacks on information systems, in particular, directly affect the operational processes that support businesses, potentially leading to corporate paralysis. Some industries are at more risk than others due to their highly sensitive data, one of which is the transportation industry, which has recently moved from traditional data management to digitalization. This transition has raised concerns about cybersecurity and necessitated proper risk assessments due to their importance in protecting critical infrastructure; for instance, cyberattacks on aircraft, which are considered essential transportation, can impact safety-of-flight systems and/or the systems supporting the airlines' business [3]. Cyber threats often take advantage of the increased complexity of infrastructure systems, placing critical industries' security at risk [4]. A physical cyber threat not only harms the integrity of the IPs but may also disrupt production processes and cause serious damage to various systems [5]. To

understand cyberattacks, it is important to dig deeply and identify their main causes. Spreading awareness and proper knowledge about cyberattacks and providing sufficient training can reduce the damage they cause. This is often difficult to accomplish because cybersecurity behaviors do not necessarily come naturally, and people need support and encouragement to develop and adopt them [1]. As technology becomes increasingly present in daily life, cybercrime, and cybersecurity tools and techniques require innovative solutions at all organizational levels [4].

Transportation systems, in particular, offer major services that can be put at risk by an absence of real awareness, and neglecting the proper assessment of vulnerabilities can lead to major damage [6]. Cyberattacks on transportation technologies are usually unexpected and require considerable effort to classify the threats, identify impacted assets, develop proper countermeasures, and engage IT teams throughout the process. However, transportation systems vary in their ability to handle threats and in the ways in which organizations prioritize their assets when a risk is identified. This paper discusses how risks to booking systems in the transportation industry are assessed at times of risk and presents a comprehensive cybersecurity risk assessment of information systems in a transportation company to identify the most common threats and recommend methods for minimizing risks as much as possible. A risk assessment report was prepared by industry experts that included previous cyberattack scenarios. This paper aims to answer the following questions:

What are the common types of cyberattacks on transportation systems?

What are the main techniques used to identify vulnerabilities in transportation systems?

What are the main risks and countermeasures used to mitigate these risks?

### B. Motivation

Understanding the nature of cyberattacks and their main causes can enhance the overall cybersecurity of an organization. A cyber threat may disrupt production processes and cause serious damage to various systems [5]. Identifying the root cause of such problems can help organizations solve them at a deep level and avoid future attacks rather than relying on temporary prevention solutions. Information systems generally contain critical data that businesses place a high priority on protecting. Some industries, such as the

transportation industry, hold more sensitive data than others; hence, their risks from cyberattacks are huge and can directly impact operational processes. It is therefore vital for them to identify the main causes of cyberattacks and the main practices they should adopt to protect sensitive data from exposure. Cybersecurity for transportation systems has been affected by the dynamic nature of the technology used within the industry. Cybersecurity guidelines have been developed for transportation systems, especially in the past few years, to ensure cybersecurity and raise awareness of its importance [6].

### C. Cybersecurity

The main reasons for cybersecurity failures are human error and insufficient knowledge [1]. According to [2], cybersecurity is central to all technologies, standards, and procedures developed to protect infrastructure elements against serious cyberattacks. Some cyberattacks cause major harm to system users, sometimes unintentionally [7]. In other words, cybersecurity protects property rights in an infrastructure context if an attack occurs. Furthermore, cybersecurity is concerned with related issues such as access, extraction, manipulation, or modification of property [8], protecting property against the harm that can be caused by an attack [7]. To maintain a secure environment, effective cybersecurity behaviors must be identified and promoted to raise awareness among users from different backgrounds. Both human and technological aspects of information systems need to be clearly identified to maintain a strong cybersecurity environment [1].

1) *Cybersecurity in information systems:* Today’s technology allows for easy, rapid communication across different systems, particularly in domains such as teleworking and m-commerce, which have grown rapidly [9]. Moreover, information and communication technology (ICT) applications have increased dramatically and cyberattacks have spread easily across such applications [10]. The more sensitive the data is, the greater attention needs to be paid. Sensitive data can be vital for businesses because they use it to make critical decisions; major problems can result from cyberattacks that place data at risk of exposure. Protecting infrastructure is a major priority for preventing unauthorized access that can lead to data misuse or corruption. Both individuals and organizations can suffer hugely from data exposure [11].

Recently, cyberattacks have increased due to advances in the technologies used in most information systems. Consequently, most organizations need to invest in cybersecurity and employee training to raise awareness of the importance of securing systems and their sensitive information [12]. One approach to protecting information systems was suggested by [13], which suggested that integrating information systems across organizational environments can improve cybersecurity. The researchers suggested and tested three hypotheses to investigate whether integration is positively related to cybersecurity countermeasures (see Table I).

Although [14] suggested considering all ICS features, the researchers proposed a targeted multilevel Bayesian network for identifying attacks, the functional level of attacks, and

incident models. This dynamic cybersecurity risk assessment approach can help assess the risks caused by unknown attacks (see Fig. 1).

Study [10] evaluated power supply reliability using Stackelberg Security Game (SSG) strategies to assign defense resources to various cyber-threat targets. This paper discussed how to benefit from the intrusion tolerance capability of SCADA systems that provide buffer periods before the failure of substations. The overall goal was to improve network strength in the face of cyber threat events. Different cyber threat scenarios were tested to assess intrusion tolerance capabilities, and the authors designed an insurance premium principle to provide incentives for enhancing intrusion tolerance capability.

Study [5] conducted a literature review to identify the impact of cyberattacks on total productive maintenance in smart manufacturing systems. Cyberattacks can directly affect manufacturing equipment and, hence, the services provided, including maintenance services. This paper highlighted major physical cyberattacks and proposed countermeasures to reduce the negative impact of such attacks. The authors identified different challenges in enhancing equipment effectiveness in light of current cybersecurity threats in the manufacturing industry.

TABLE I. HYPOTHESES AND EVIDENCE SUMMARY [12]

| Hypotheses                                                                                                                            | Findings  | Evidence                                                                                                                                                                                |
|---------------------------------------------------------------------------------------------------------------------------------------|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| H1. The greater the integration of IS, the greater the investment in countermeasures.                                                 | Supported | IS integration causes fewer weak points, reducing the possible impact of breaks.                                                                                                        |
| H2. H1 will be more powerful when considering external IS integration rather than internal IS integration.                            | Supported | Weak points in external IS integration involve greater risk exposure because of greater uncertainty.                                                                                    |
| H3. Organizations tend to use self-protective controls more often in highly volatile environments than in less volatile environments. | Supported | Although the impact may not be strong, volatile environments can impact the three aspects of vulnerability. This means that the addressing of weak points must highlight these aspects. |

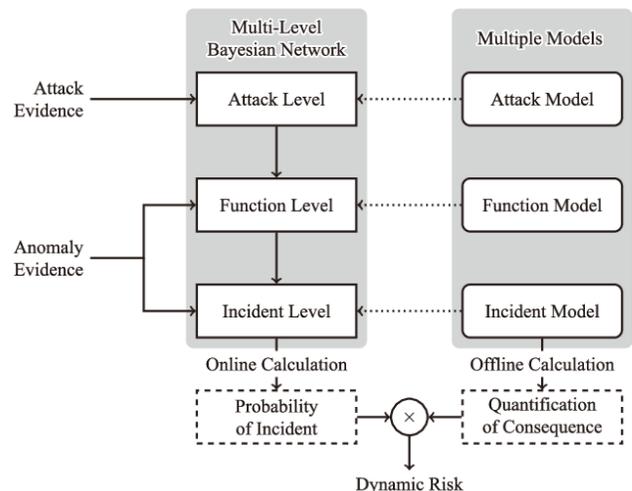


Fig. 1. Architecture of the Dynamic ICS Cybersecurity Approach.

According to study [15], attack graphs are essential for identifying the variables involved in an attack and reducing their impact on networks. This research introduced a cyberattack path method that used restrictions and an in-depth search to successfully produce attack graphs according to the interests of users. The researchers used real data from a maritime supply chain to ensure the validity of the proposed method.

In [16], the author identified the effects of cyberattacks on general systems. As cyberattacks continue to develop, it is becoming more difficult to identify the nature of the attacks; therefore, there is a great need for smart risk assessment. This research proposed the use of a fuzzy inference (FIS) model to produce risk assessment outputs, which relied on four risk factors—vulnerability, threat, likelihood, and impact—to identify risks targeting a system entity and suggest possible solutions for them. A summary of related work is provided in Table II.

2) *Cybersecurity threats in the transportation industry:* The transportation industry needs to distinguish between operations systems and business systems to provide the right protection for each [6]. Over the years, the industry has shifted from traditional business to e-business, and this shift has expanded technologies and their features [2]. According to [11], 80 % of assets in transportation infrastructure are being digitalized. In recent years, many attacks have been made on transportation, which has increased the need for cybersecurity protection guidelines [6], and some factors are critical for ensuring the effectiveness of overall cybersecurity, such as PCS systems, knowledge about cyber threats, and communication between private corporations and public agencies [17]. In the air transportation domain, cybersecurity tends to focus greatly on protecting the operational and technical aspects of businesses; hence, fast adaption to a rapidly changing risk environment is vital, and the framework of technical and operational systems should be redesigned based on continuous risk analysis and simulations [18]. The rapidly changing nature of the transportation industry makes it important to focus on cybersecurity to protect valuable assets and protect the business from harmful threats.

Study [18] was conducted to address the increase in cyberattacks, the impact of which could critically affect civil

aviation functions. The huge increase in technologies and integrated connectivity tools can expose air traffic management (ATM) to major risk, despite its high value as an asset. This study evaluated cybersecurity difficulties in ATM to develop a threat model that included likely risks. It also included an overall framework that required full collaboration between entities to identify threats and protect systems from attacks.

Study [19] asserted that the port industry is experiencing a transformation in connectivity between ports, where most functions are being digitalized. This necessitates focusing on cybersecurity to protect major infrastructure against advanced attacks and maximize the use of new technologies with minimum risk of affecting valuable business assets.

Study [11] highlighted the importance of data-driven functions that many business aspects depend on, such as operations, maintenance, planning, and decision-making. To ensure the smooth operation of all functions relating to railways, data should be strongly secured against cyberattacks and unauthorized access to avoid major losses. This paper identified possible challenges, impacts, threats, vulnerabilities, and methods for managing risks and protecting railway infrastructure data, particularly in an e-maintenance context.

Study [6] used a case study to raise awareness of the cybersecurity attacks that affect the transportation field. It developed an attack–fault tree for the mentioned case study as proof of concept for integrated risk analysis. The overall purpose was to help companies understand that no attacks targeting critical technological systems should be ignored, and potential risks should be analyzed.

The author in [3] proposed a new system for gathering, managing, and reporting aircraft failures. The motivation behind this paper was the great expansion in connectivity and communication infrastructure that is affecting aircraft. The increase in mobile computing device use among individuals has allowed for external connectivity increments as well as providing internet access for passengers, involving a greater risk of aircraft cyberattacks that can affect other critical systems supporting the business. The proposed system can help identify such attacks, hence reducing their impact. A summary of related work in the transportation domain is provided in Table III.

TABLE II. SUMMARY OF RELATED WORK

| Study  | A Cybersecurity Insurance Model for Power System Reliability Considering Optimal Defense Resource Allocation [10] | Multimodal-Based Incident Prediction and Risk Assessment in Dynamic Cybersecurity Protection for Industrial Control Systems [14] | Cybersecurity Concerns for Total Productive Maintenance in Smart Manufacturing Systems [5] | Improving Risk Assessment Models of Cyber Security Using a Fuzzy Logic Inference System [16] | Cyberattack Path Discovery in a Dynamic Supply Chain Maritime Risk Management System [15] |
|--------|-------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------|
| Domain | Cyber physical systems (CPSs)                                                                                     | Industrial control systems (ICSSs).                                                                                              | Manufacturing systems                                                                      | General system                                                                               | Dynamic supply chain maritime risk management system                                      |
| System | Modern power grids                                                                                                | Simplified chemical reactor control system                                                                                       | Total productive maintenance (TPM)                                                         | Various system entities                                                                      | Maritime supply chain                                                                     |

|                                                                 |                                                                                                                                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|-----------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <p><b>Purpose</b></p>                                           | <p>To benefit from the intrusion tolerance capability of SCADA systems that provide buffer periods before the failure of substations. The overall goal was to improve network strength and counter cyber threats.</p>                                 | <p>To develop a dynamic risk assessment approach that could identify risks due to unknown threats and enhance the accuracy of risk assessment processes.</p>                                                                                                                                                                                   | <p>To illustrate the impact of cyberattacks on total productive maintenance in smart manufacturing systems and to discuss countermeasures to reduce the negative impact of an attack.</p>                                                                                                                                                                                                                                                                                                                                                                         | <p>To use a fuzzy inference (FIS) model to produce risk assessment outputs, which relied on four risk factors—vulnerability, threat, likelihood, and impact—to identify risks targeting a system entity and suggest possible solutions for such threats.</p>                                                                                                                                                                                                                                                                                                | <p>To introduce a cyberattack path method that used restrictions and an in depth search to successfully produce attack graphs according to the interests of users using real data from a maritime supply chain to ensure the validity of the proposed method.</p>                                                                                                                                                                                                    |
| <p><b>Possible Threats</b></p>                                  | <ul style="list-style-type: none"> <li>• a denial-of-service (DoS) attack</li> <li>• bypassing the VPN to gain access to the servers</li> <li>• changes in voltage and standard measurements</li> </ul>                                               | <ul style="list-style-type: none"> <li>• malicious attacks</li> <li>• spoof attacks</li> <li>• breaches of an intrusion detection system (IDS)</li> </ul>                                                                                                                                                                                      | <ul style="list-style-type: none"> <li>• intellectual properties threats, including theft and data modification</li> <li>• cyberphysical threats that disrupt production processes</li> <li>• a Stuxnet worm infection</li> <li>• malicious void attacks</li> </ul>                                                                                                                                                                                                                                                                                               | <ul style="list-style-type: none"> <li>• website attacks</li> <li>• malware</li> <li>• hacking</li> <li>• denial of service (DoS)</li> <li>• name hijackings</li> <li>• dissemination of viruses.</li> <li>• phishing and spam e-mails</li> </ul>                                                                                                                                                                                                                                                                                                           | <p>Attack paths within a network:</p> <ul style="list-style-type: none"> <li>• DoS attacks</li> <li>• distributed denial of service (DDoS) attacks</li> </ul>                                                                                                                                                                                                                                                                                                        |
| <p><b>Risk assessment enhancement (previous approaches)</b></p> | <ul style="list-style-type: none"> <li>▪ Component burnout and exhaustion of processing power.</li> <li>▪ Simulating and forecasting real-time load to ensure system frequency during an attack.</li> </ul>                                           | <ul style="list-style-type: none"> <li>▪ IDS to observe network and system activities.</li> <li>▪ An anomaly detection system (ADS) to gather data from a system and compare them with normal values (reports produced in cases of deviation)..</li> </ul>                                                                                     | <ul style="list-style-type: none"> <li>▪ Use of overall equipment effectiveness (OEE), which is considered a major KPI for measuring the effectiveness of TPM in a system. The OEE of a system is calculated using the input of three components: <ul style="list-style-type: none"> <li>- breakdowns (availability)</li> <li>- small stops (performance)</li> <li>- defects (quality)</li> </ul> </li> <li>▪ Each component can be impacted by a cyberattack.</li> <li>▪ <math>OEE = \text{availability} * \text{performance} * \text{quality}</math></li> </ul> | <ul style="list-style-type: none"> <li>▪ To deal with the uncertainty factor when gathering data, the fuzzy set theory can help in making decisions about various alternatives. Despite its ability to deal with fuzziness, only a few studies have used fuzzy set theory to handle risk uncertainty, although it is highly recommended for improving the use of this theory for critical risk assessment.</li> <li>▪ Existing models without human intervention.</li> </ul>                                                                                | <ul style="list-style-type: none"> <li>▪ MulVal network security analyzer to target bugs within network configurations.</li> <li>▪ TVA tool for topological network-based analysis.</li> <li>▪ A graph model based on a specific language to simulate attack scenarios using various methods.</li> <li>▪ An intrusion detection system to generate graphs for attacks.</li> <li>▪ NuSMV model for allocating vulnerabilities and producing attack graphs.</li> </ul> |
| <p><b>Proposed Contribution/ Recommendation</b></p>             | <p>A Stackelberg Security game model to allocate defense resources, unknown to the attacker. Encouraging investment in defense resource coverage to improve the intrusion tolerance capability of SCADA systems and protect them against failure.</p> | <p>The proposed solution is capable of measuring cybersecurity risks of ICSs in a A short-term multimodal-based cybersecurity risk assessment approach with the ability to produce cybersecurity risk values by calculating the probabilities of risks and quantifying the impacts of different possible incidents caused by cyberattacks.</p> | <p>Acquiring an agile maintenance system and considering both mean time between failures (MTBF) and mean time to repair (MTTR), relying on a short repair time. A proposed plan for system recovery, enabling repairs to be performed as quickly as possible.</p>                                                                                                                                                                                                                                                                                                 | <p>The proposed solution senses a weak item and moves it to a risk assessment model, which then determines the items for the spatial computation methods and passes them to the next model for approval. Approval suggests the end of the process. However, if an item is not approved, it will be moved to other models for vulnerability estimation using fuzzy theory. Information will be displayed to interested parties, enabling them to decide mitigating actions. The process starts again, relying on human judgment to decrease uncertainty.</p> | <ul style="list-style-type: none"> <li>• The proposed method identifies specific paths in a certain network to enhance risk assessment. These paths are unique, such as: <ul style="list-style-type: none"> <li>○ attacker capability</li> <li>○ attacker location</li> <li>○ propagation length</li> <li>○ maximum length</li> <li>○ entry points</li> <li>○ target points</li> </ul> </li> </ul>                                                                   |

TABLE III. SUMMARY OF RELATED WORK IN THE TRANSPORTATION DOMAIN

| Study                                        | Aviation Cybersecurity and Cyber-Resilience: Assessing Risk in Air Traffic Management [18]                                                                                                                                                                                                                                                                                                                                                                                                                     | Cybersecurity in Ports and the Maritime Industry: Reasons for Raising Awareness on This Issue [19]                                                                                                                                                                                                                                                                                                        | Cybersecurity for eMaintenance in Railway Infrastructure: Risks and Consequences [11]                                                                                                                                                                                                                                                                                                                                                                                                                                                     | Cybersecurity and its Integration with Safety for Transport Systems: Not a Formal Fulfillment but an Actual Commitment [6]                                                                                                      | A System for Real-time Monitoring of Cybersecurity Events on Aircraft [3]                                                                                                                                                                                                             |
|----------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Domain</b>                                | Air transportation                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | Port industry                                                                                                                                                                                                                                                                                                                                                                                             | Railway industry                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          | Railway industry                                                                                                                                                                                                                | Air transportation                                                                                                                                                                                                                                                                    |
| <b>System</b>                                | ATM                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | Port 4.0                                                                                                                                                                                                                                                                                                                                                                                                  | E-maintenance in railway infrastructure                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   | Subsystem for railway vehicles (wheel slide protection [WSP])                                                                                                                                                                   | Aircraft                                                                                                                                                                                                                                                                              |
| <b>Purpose</b>                               | To analyze potential targets and risks.                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | To maximize the benefits of using full technology while ensuring that major infrastructure elements are well protected against cyberattacks.                                                                                                                                                                                                                                                              | To identify possible difficulties, impacts, and risks of data security for railway infrastructure, and to highlight methodologies for attaining and securing data against possible breaches.                                                                                                                                                                                                                                                                                                                                              | To enhance awareness of possible weaknesses that impact transport systems. Also, to install spotting lights on the embedded devices used by those systems and prevent major attacks that can target them if not well protected. | To track and monitor incidents/failures and protect aircraft and related systems from cyberattacks.                                                                                                                                                                                   |
| <b>Possible Threats</b>                      | <ul style="list-style-type: none"> <li>passive observers</li> <li>activists and lobbyists</li> <li>insiders</li> <li>cyber crime</li> <li>cyber terrorism</li> <li>hostile nation-states</li> </ul>                                                                                                                                                                                                                                                                                                            | <ul style="list-style-type: none"> <li>organized criminal rings</li> <li>drug traffickers</li> <li>terrorists</li> <li>hackers</li> <li>industrial spies and competitors</li> <li>disgruntled staff and insiders, enemy states, and foreign intelligence</li> </ul>                                                                                                                                       | <ul style="list-style-type: none"> <li>data theft</li> <li>database breaches</li> <li>targeting of application servers</li> <li>stealing of authentication details from system administrators</li> <li>data integrity being affected by modification actions</li> <li>DDoS</li> <li>directed denial of service attacks</li> <li>physical annihilation attacks</li> </ul>                                                                                                                                                                  | <ul style="list-style-type: none"> <li>physical attacks: installing malicious devices</li> <li>side-channel attacks to obtain encryption keys</li> <li>logical attacks: malicious code injections</li> </ul>                    | <ul style="list-style-type: none"> <li>delayed aircraft flight operations</li> <li>compromised safety of flight systems</li> <li>high recovery costs affecting business</li> <li>theft of passengers' personal data.</li> <li>malware deployed on multiple targets.</li> </ul>        |
| <b>Current Security Measures</b>             | <ul style="list-style-type: none"> <li>physical security (e.g., access control)</li> <li>personnel security (e.g., security clearances)</li> <li>information security (e.g., software updates and patches)</li> <li>communication security (e.g., network segregation)</li> <li>intelligence support (e.g., security alert level declarations)</li> <li>security information exchanges (e.g., incident identification and notification)</li> <li>operational continuity (e.g., emergency responses)</li> </ul> | <ul style="list-style-type: none"> <li>Increase awareness among port ecosystem parties by:</li> <li>publishing standards to address cybersecurity issues</li> <li>issuing shipping company guidelines and recommendations</li> <li>requesting the inclusion of cybersecurity in facility security assessments to address any vulnerabilities</li> <li>publishing a Guide on Port Cybersecurity</li> </ul> | <ul style="list-style-type: none"> <li>General examples:</li> <li>inventory of devices/software</li> <li>malware defenses</li> <li>application software security</li> <li>wireless device control</li> <li>data recovery capability</li> <li>security skill assessments and training</li> <li>protection of network ports and services</li> <li>boundary defense</li> <li>security audit logs</li> <li>account monitoring and control</li> <li>data loss prevention</li> <li>incident response capability</li> <li>penetration</li> </ul> | <ul style="list-style-type: none"> <li>Current strategy of risk assessment is based on single threats and compliance to specific practices, leading to neglect of the effects of combined hazardous events.</li> </ul>          | <ul style="list-style-type: none"> <li>Current systems include logging and monitoring of failures as maintenance data. This approach does not allow prompt tracking of security attacks on aircraft networks, which can allow successful attacks with no detectable trace.</li> </ul> |
| <b>Proposed Contribution/ Recommendation</b> | This study proposed an interactive and model-based cyber risk analysis                                                                                                                                                                                                                                                                                                                                                                                                                                         | Policymakers should work closely with industry to ensure full                                                                                                                                                                                                                                                                                                                                             | <ul style="list-style-type: none"> <li>Enhancing confidentiality: ensuring data privacy (i.e.,</li> </ul>                                                                                                                                                                                                                                                                                                                                                                                                                                 | An integrated safety and cybersecurity analysis of all                                                                                                                                                                          | All apps should send security event failure logs for security                                                                                                                                                                                                                         |

|  |                                                                                 |                                                                                                                                                                                                                                                                                    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                               |                                                                                                                                                                                                                                                                                                                                                                                                                                     |
|--|---------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | that could produce non-stop cyber flexibility in the air transportation domain. | protection of multiple port systems because they have a major impact on the global economy. Also, they should continuously review current policies and regulations and adopt new industrial technologies. Moreover, they should invest in alert systems to detect cyber incidents. | targeted data accessed/viewed only by authorized individuals). <ul style="list-style-type: none"> <li>Enhancing integrity: <ol style="list-style-type: none"> <li>Supporting data authenticity by using digital signatures or other trusted identifiers.</li> <li>Avoiding data errors when transferring/storing data and making sure that data are original.</li> </ol> </li> <li>Enhancing availability: ensuring that data access is granted to authorized parties.</li> </ul> | related control systems could reduce the impact of major threats, as suggested by this study. | monitoring and assessment and: <ol style="list-style-type: none"> <li>comprise similar applications</li> <li>capture security event failure logs from applications and services on aircraft</li> <li>manage the logs for essential security event failures</li> <li>alert crew for fast recovery and communication with ground system in case of major failures</li> <li>maintain the logs for future maintenance usage.</li> </ol> |
|--|---------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## II. CASE STUDY: RISK ASSESSMENT

### A. Scenario

Daily DDoS attacks against company systems are a great concern for IT managers; however, the previous severe DDoS attack, which was repeated twice, resulted in approximately four hours of total downtime, was extremely intense, and aimed to fully disrupt the company's booking services, which could have had a significant financial impact. IT leaders directed the cybersecurity team to immediately conduct a risk assessment of these cyberattacks and provide feedback for decision-making. A risk analysis report was prepared using various cybersecurity risk management methodologies to overcome the above-mentioned issues, and the general scenario related to "the risk associated with cyberattacks against the availability of the booking system." [22-26].

### B. Risk Assessment

The company follows a combined approach to risk assessment, which is managed by the Cybersecurity Department and the IT Governance, Risk, & Audit (GRA) Department. Their goal is to ensure the management of information technology and security risks [27-32].

1) *Asset identification*: To identify the assets related to the system, system functions were first had identified [32-38]. The scope of the risk assessment was the company's booking system, represented by an application that provides reservation and ticketing services to various transport sectors through the company's digital channels (see Table IV). List of the most common risks and their corresponding controls targeting booking systems is shown in Table V.

2) *Threat and vulnerability identification*: Table VI contains the most common threat types targeting web-based systems and their threat communities. Due to the high level of data sensitivity, vulnerabilities were derived from study [20], which highlighted the most common vulnerabilities of Web-based systems but did not necessarily reflect the actual company's data [39-40].

Vulnerabilities can be divided into two classes. The first class includes vulnerabilities that affect a host or only a service running on it:

- host crash.
- performance fault.
- host infection.

The second class includes vulnerabilities that affect only a single service:

- inaccessible service.
- corrupted service.

3) *Techniques to identify vulnerabilities*: Companies use various techniques to identify vulnerabilities in their systems, and this paper identifies the set of techniques used by transportation companies; for instance, the network security team scans the system a number of times daily, and firewalls and scanners are in place to detect spikes in incoming traffic. Additionally, a DDoS protection service is in place to protect the system. The IT Security team conducts regular exercises to identify vulnerabilities using various technologies, including system vulnerability scans, penetration testing, Web application assessments, and network mapping. Furthermore, the IT team conducts special system scans for indicators of compromise upon requests from the NCA. The company also has monitoring, incident response, and forensics teams working closely with security business partners to cover various areas, such as system logs and audit reports.

### C. Minimizing Risks

The chosen risk was based on the two previous high-DDoS incidents that affected the transportation company's system. Management direction played a critical role in selecting what type of risk to manage (see Table VII).

1) *Threat community profile*: Each threat was known to have its own community profile and could have different initiating factors or triggers. Below are common factors relating to cybersecurity attacks (particularly regarding DDoS; see Table VIII).

TABLE IV. ASSET IDENTIFICATION AND CORRESPONDING VALUES

| System functions                                                                                                                                                                                                                                                                                                                                      | System elements                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        | Related department                                                                                                                                                                                                     | Number of employees | Assets                                                                                                                                                                                                       | Value                                                                                                                                                                                                                                                                                                                        |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ol style="list-style-type: none"> <li>1. booking of tickets for trips, cars, trains, hotels, etc.</li> <li>2. lounge access</li> <li>3. requests for trip upgrades</li> <li>4. loyalty programs</li> <li>5. online payments</li> <li>6. online check-ins</li> <li>7. service refunds</li> <li>8. real-time trip information and schedules</li> </ol> | <ol style="list-style-type: none"> <li>A. <b>Input:</b> <ol style="list-style-type: none"> <li>1. trip schedule</li> <li>2. locations</li> <li>3. customer information</li> </ol> </li> <li>B. <b>Processing:</b> <ol style="list-style-type: none"> <li>1. booking of trips</li> <li>2. payment</li> <li>3. checking in</li> </ol> </li> <li>C. <b>Output:</b> <ol style="list-style-type: none"> <li>1. scheduled trips</li> <li>2. booking reservations</li> <li>3. ticket passes</li> <li>4. marketing campaigns</li> </ol> </li> <li>D. <b>Interface:</b> <ol style="list-style-type: none"> <li>1. website</li> <li>2. mobile application</li> </ol> </li> </ol> | <ol style="list-style-type: none"> <li>1. <b>Business:</b> marketing and ticketing services</li> <li>2. <b>IT:</b> digital products and services</li> <li>3. <b>Others:</b> vendor and IT business partners</li> </ol> | 20 employees        | <ul style="list-style-type: none"> <li>• servers</li> <li>• firewalls</li> <li>• databases</li> <li>• micro services</li> <li>• application gateway</li> <li>• VPN gateway</li> <li>• API gateway</li> </ul> | <p><b>Information:</b> customers' data, such as national IDs and credit-cards, are considered the mostvaluable asset in this system.</p> <p><b>Internal HW/SW:</b> support that helps with various functions of the system.</p> <p><b>Vendor Services:</b> security services that protect against availability attacks).</p> |

TABLE V. A LIST OF THE MOST COMMON RISKS AND THEIR CORRESPONDING CONTROLS TARGETING BOOKING SYSTEMS

| Risk                                                      | Counter Measures                                                                                                                          |
|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------|
| Suspected phishing domain similar to the company website. | <ul style="list-style-type: none"> <li>• block the domain.</li> <li>• request to take down the domain</li> </ul>                          |
| A copy of a company application.                          | <ul style="list-style-type: none"> <li>• request to remove the app</li> </ul>                                                             |
| Employee login credentials on the dark Web.               | <ul style="list-style-type: none"> <li>• check the accounts</li> <li>• reset passwords</li> <li>• enable MFA</li> </ul>                   |
| Company internal environment exposed.                     | <ul style="list-style-type: none"> <li>• hide the internal environment</li> <li>• restrict access to authorized personnel only</li> </ul> |
| Malware detected internally.                              | <ul style="list-style-type: none"> <li>• remove the malware</li> </ul>                                                                    |
| User logon from a risky IP address.                       | <ul style="list-style-type: none"> <li>• check with the user</li> <li>• block the IP</li> </ul>                                           |
| Activity from a Tor IP address.                           | <ul style="list-style-type: none"> <li>• check with the user</li> <li>• block the IP</li> </ul>                                           |
| Files shared with unauthorized domain.                    | <ul style="list-style-type: none"> <li>• check with the user</li> <li>• block the domain</li> </ul>                                       |

TABLE VI. COMMON THREATS TARGETING WEB-BASED SYSTEMS AND THEIR COMMUNITIES

| Type                    | Threat Community (source)         | Asset at Risk  | Effect          |
|-------------------------|-----------------------------------|----------------|-----------------|
| DDoS                    | Cyber criminals                   | Booking system | Availability    |
| SQL injection           | Cyber criminals                   | Booking system | Confidentiality |
| SQL injection           | Cyber criminals                   | Booking system | Integrity       |
| Cross site scripting    | Cyber criminals                   | Booking system | Confidentiality |
| Cross site scripting    | Cyber criminals                   | Booking system | Integrity       |
| SQL injection           | Script kiddies                    | Booking system | Confidentiality |
| SQL injection           | Script kiddies                    | Booking system | Integrity       |
| Privilege escalation    | Privileged insiders and employees | Booking system | Confidentiality |
| Privilege escalation    | Privileged insiders and employees | Booking system | Availability    |
| Privilege escalation    | Privileged insiders and employees | Booking system | Integrity       |
| Bad bots                | Cyber criminals                   | Booking system | Confidentiality |
| Illegal resource access | Cyber criminals                   | Booking system | Confidentiality |
| Phishing                | Social engineer                   | Booking system | Confidentiality |

TABLE VII. THE SELECTED RISK

| Asset at Risk  | Threat Community | Type | Effect       |
|----------------|------------------|------|--------------|
| Booking system | Cyber criminals  | DDoS | Availability |

TABLE VIII. DDoS COMMUNITY PROFILE

| Factor                                   | Value                                           |
|------------------------------------------|-------------------------------------------------|
| Motive                                   | Financial disruption.                           |
| Primary intent                           | Illegal activities to maximize profit.          |
| Sponsorship                              | Non-state or illegal gangs.                     |
| Preferred general target characteristics | Easy financial gains via remote means.          |
| Preferred targets                        | Financial services and retail organizations.    |
| Capability                               | Professional, skilled, and well-funded hackers. |
| Personal risk tolerance                  | Relatively high, without being exposed.         |
| Concern for collateral damage            | Prefer to keep their identities hidden.         |

D. Likelihood Estimation

Threat event frequency (TEF) was used to estimate the likelihood of a threat, indicating the probable frequency within a given timeframe that a threat would result in loss (see Table IX).

TABLE IX. THREAT EVENT FREQUENCY (TEF) FOR A DDoS ATTACK

| TCom            | Threat Type | TEF Min                     | TEF ML                        | TEF Max                         |
|-----------------|-------------|-----------------------------|-------------------------------|---------------------------------|
| Cyber criminals | DDoS        | 365 (per year)<br>1 (daily) | 1,825 (per year)<br>5 (daily) | 10,220 (per year)<br>28 (daily) |

TCom: Threat community (source)

TEF Min: Minimum threat event frequency (attack frequency)

TEF ML: Most likely threat event frequency (attack frequency)

TEF Max: Maximum threat event frequency (attack frequency)

Similarly, loss-even frequency (LEF) was calculated to indicate the probable frequency within a given timeframe of a loss being expected to occur (see Table X).

TABLE X. LOSS EVENT FREQUENCY (LEF) FOR A DDoS ATTACK

| TCom            | Threat Type | LEF Min    | LEF ML       | LEF Max      |
|-----------------|-------------|------------|--------------|--------------|
| Cyber criminals | DDoS        | 1 per year | 2 (per year) | 4 (per year) |

1) *Likelihood scale for the identified risk:* According to the previously identified incident, the likelihood of a DDoS attack being successful was 2 (as per the previous incident). Table XI was used to derive the loss event frequency (likelihood) and total risk category to be input into the risk matrix.

2) *Impact identification:* The table below shows the total impacts due to loss of availability. Impact types varied between lost revenue, the cost of hiring an incident response team, and the cost of investigating the crime (i.e., forensics cost; see Table XII).

Table XIII shows the availability impact scale used by the company to identify the severity of an impact for the risk matrix.

TABLE XI. LIKELIHOOD SCALE FOR DDoS RISK

| Score | Rating                   | X | Description                          |
|-------|--------------------------|---|--------------------------------------|
| 4     | Very high (VH)           |   | More than 5 likelihood of occurrence |
| 3     | High (H)                 |   | 4–5 likelihood of occurrence         |
| 2     | Medium (M)               | X | 2–3 likelihood of occurrence         |
| 1     | Very low to unlikely (L) |   | 0–1 likelihood of occurrence         |

TABLE XII. TOTAL IMPACT

| Impact Type                      | Min. (1–2 h downtime) | Most Likely (3–5 h downtime)     | Max. (10 h downtime) |
|----------------------------------|-----------------------|----------------------------------|----------------------|
| Lost revenue                     | 1,050,000             | 2,625,000                        | 5,250,000            |
| Incident response team(internal) | 5,000                 | 7,800                            | 10,000               |
| Forensics (external)             | 50,000                | 56,250                           | 60,000               |
| Total                            | 1,105,000             | 2,689,050 (rounded)<br>2,700,000 | 5,320,000            |

TABLE XIII. AVAILABILITY IMPACT SCALE

| Risk Rating | Impact                                                                                                                                                                                                       |
|-------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Low         | <ul style="list-style-type: none"> <li>no significant effect on operations and services</li> <li>asset can be replaced within an acceptable time frame</li> <li>insignificant interruption costs</li> </ul>  |
| Medium      | <ul style="list-style-type: none"> <li>no significant effect on operations and services</li> <li>asset can be replaced within a medium time frame</li> <li>low interruption costs</li> </ul>                 |
| High        | <ul style="list-style-type: none"> <li>effect on individual operations and services</li> <li>critical assets cannot be replaced by manual methods</li> <li>high interruption costs</li> </ul>                |
| Very High   | <ul style="list-style-type: none"> <li>significantly affects multiple operations and services</li> <li>critical assets cannot be replaced by manual methods</li> <li>very high interruption costs</li> </ul> |

According to the scenario provided by the company’s IT team, the DDoS attack was repeated twice, resulting in an approximate downtime of four hours (see Table XIV).

3) *Risk matrix:* The following risk matrix includes two factors: impact and likelihood. Both factors have a rating scale of 1–4, as shown in the previous scaling tables. The IT team identified the likelihood of the risk occurring as stated in the scenario (i.e., twice a year; medium rating = 2), and the teams also measured the loss impact of four hours of total system downtime (very high rating = 4). The risk level was then calculated as the likelihood of risk occurrence \* impact of a loss, resulting in a risk level of eight (see Table XV).

The company’s main risk objective was to protect the organization’s information and technology assets by maintaining confidentiality, integrity, and availability of service effectively with minimum cost and without affecting business operations. The strategy for responding to risks

depended on the individual risk situation and was based on risk assessments and recommendations from decision-makers. As shown in Table XV, the risk level was relatively high and needed to be managed; hence, the transportation company decided to mitigate the risk by applying appropriate countermeasures. A list of countermeasures suggested by IT experts was prepared by the transportation company's IT team (see Fig. 2).

a) Internal controls

Procedures: enhance the DDoS Response Plan with:

- a systems checklist including all assets to ensure advanced threat identification and assessment.
- notifications and escalation procedures for quick recovery.

Training:

- train special teams to extensively monitor traffic and look for abnormalities, including unexplained traffic spikes and visits from suspect IP addresses and geolocations.
- create additional response teams to minimize the impact of attacks.

TABLE XIV. IMPACT SCALE FOR A DDoS ATTACK

| Score | Rating    | X | Description               |
|-------|-----------|---|---------------------------|
| 4     | Very high | X | More than 3 h downtime    |
| 3     | High      |   | 1–3 h downtime            |
| 2     | Medium    |   | 30 min–1 h downtime       |
| 1     | Low       |   | Less than 30 min downtime |

TABLE XV. DDoS RISK MATRIX

|            |   |   |   |    |    |
|------------|---|---|---|----|----|
| Likelihood | 4 | 4 | 8 | 12 | 16 |
|            | 3 | 3 | 6 | 9  | 12 |
|            | 2 | 2 | 4 | 6  | 8  |
|            | 1 | 1 | 2 | 4  | 5  |

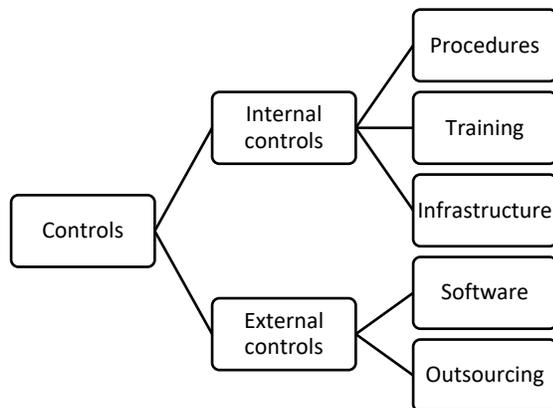


Fig. 2. List of Suggested Countermeasures.

Infrastructure:

- create redundant network (FW/IPS) resources so that, if one server is attacked, the others can handle extra network traffic.

b) External controls

Software:

- purchase threat intelligence software to monitor social media and the dark Web for threats, suspicious conversations, and boasts that may hint at an incoming attack.

Outsourcing:

- use third-party DDoS testing (i.e., pen testing) to simulate attacks against IT infrastructure so that the company can be prepared for any real threats.
- Use DDoS-as-a-service to provide improved flexibility for environments that combine in-house and third-party resources, or cloud and dedicated server hosting.
- outsource DDoS prevention to cloud-based service providers operated by software engineers whose job consists of monitoring the Web for the latest DDoS tactics. For decision-makers to choose between countermeasures for mitigating DDoS attacks, IT experts used the following scale to evaluate the effectiveness of each control.

Each control had a corresponding estimated cost and effectiveness rating (see Table XVI and XVII). The following criteria were used to choose the appropriate controls:

- If the control will reduce the risk more than needed, a less expensive alternative should be used.
- If the control will cost more than the risk reduction provided, an alternative should be used.
- If the control does not sufficiently reduce the risk, either more or different controls should be used.
- If the control provides sufficient risk reduction and is the most cost-effective option, use it.

TABLE XVI. CONTROL EFFECTIVENESS SCALE

| Risk Rating             | Impact                                                                                                                                                                                    |
|-------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Ineffective             | <ul style="list-style-type: none"> <li>• poor control design</li> <li>• significant control gaps</li> <li>• does not treat root causes</li> <li>• does not operate effectively</li> </ul> |
| Partially effective     | <ul style="list-style-type: none"> <li>• satisfies control design needs</li> <li>• partially treats the root causes of the risk</li> <li>• not very effective</li> </ul>                  |
| Substantially effective | <ul style="list-style-type: none"> <li>• designed correctly</li> <li>• treats most of the root causes of the risk</li> <li>• requires improvements to operate effectively</li> </ul>      |
| Fully effective         | <ul style="list-style-type: none"> <li>• well designed</li> <li>• addresses and treats all root causes</li> <li>• effective and reliable at all times</li> </ul>                          |

TABLE XVII. ESTIMATED COST FOR EACH CONTROL

| # | Control                                            | Estimated Cost | Effectiveness           |
|---|----------------------------------------------------|----------------|-------------------------|
| 1 | Enhance the DDoS response plan                     | 10,000         | Ineffective             |
| 2 | Train special teams                                | 30,000         | Partially effective     |
| 3 | Create additional response teams                   | 45,000         | Partially effective     |
| 4 | Use third-party DDoS testing                       | 75,000         | Substantially effective |
| 5 | Purchase threat intelligence software              | 100,000        | Substantially effective |
| 6 | Create redundant network resources                 | 200,000        | Substantially effective |
| 7 | DDoS-as-a-service provision                        | 350,000        | Fully effective         |
| 8 | Outsource DDoS prevention to a cloud-based service | 500,000        | Fully effective         |

c) Suggested controls for implementation: A cost-benefit analysis was conducted to identify the most appropriate controls and provide the greatest benefit to the company given the available resources. Two selected controls were recommended for implementation based on a cost-benefit analysis performed to justify why decision-makers should implement them (see Table XVIII).

TABLE XVIII. SUGGESTED CONTROLS FOR IMPLEMENTATION

| #          | Control                            | Estimated Cost | Effectiveness           |
|------------|------------------------------------|----------------|-------------------------|
| 3          | Create additional response teams   | 45,000         | Partially effective     |
| 6          | Create redundant network resources | 200,000        | Substantially effective |
| Total Cost |                                    | 245,000        | Substantially effective |

E. Cost-Benefit Analysis

The selected controls minimized the likelihood of a DDoS risk occurring twice to 0 or 1 (very low rating = 1), while the impact of DDoS was reduced from a total downtime of three hours to a medium impact (30 min–1 h), with a score of 2 (see Tables XIX and XX).

TABLE XIX. LIKELIHOOD OF A RISK AFTER IMPLEMENTING SELECTED COUNTERMEASURES

| Score | Rating                   | X | Description                  |
|-------|--------------------------|---|------------------------------|
| 4     | Very high (VH)           |   | More than 5                  |
| 3     | High (H)                 |   | 4–5 likelihood of occurrence |
| 2     | Medium (M)               |   | 2–3 likelihood of occurrence |
| 1     | Very low to unlikely (L) | X | 0–1 likelihood of occurrence |

TABLE XX. IMPACT OF THE RISK AFTER IMPLEMENTING THE SELECTED COUNTERMEASURES

| Score | Rating    | X | Description               |
|-------|-----------|---|---------------------------|
| 4     | Very high |   | More than 3 h downtime    |
| 3     | High      |   | 1-3 h downtime            |
| 2     | Medium    | X | 30 min–1 h downtime       |
| 1     | Low       |   | Less than 30 min downtime |

As shown in the risk level matrix (see Table XXI), the new risk level was calculated as the likelihood of risk occurrence \* impact of a loss, resulting in a residual risk level of two.

TABLE XXI. RESIDUAL RISK AFTER IMPLEMENTING CONTROLS

|            |   |   |    |    |
|------------|---|---|----|----|
|            | 4 | 8 | 12 | 16 |
| Likelihood | 3 | 6 | 9  | 12 |
|            | 2 | 4 | 6  | 8  |
|            | 1 | 2 | 4  | 5  |
|            |   | 1 | 2  | 3  |

III. CONCLUSION

As shown in the case study scenario, the risk assessment identified the most critical attacks on the transportation company’s booking system and provided suitable countermeasures to minimize the risk of attacks. The risk level decreased from eight to two, indicating the effectiveness of the selected countermeasures. Risk assessment was extremely useful for assessing potential risks and suggesting useful controls. Moreover, the two identified DDoS attacks were mitigated by implementing suitable controls, and recommendations were made to analyze and monitor incidents and increase the company’s preparedness for another wave of DDoS or other attacks.

REFERENCES

- [1] Y. Hong and S. Furnell, “Understanding cybersecurity behavioral habits: Insights from situational support,” *Journal of Information Security and Applications*, vol. 57, p. 102710, 2021.
- [2] Almaiah MA. A New Scheme for Detecting Malicious Attacks in Wireless Sensor Networks Based on Blockchain Technology. *Artificial Intelligence and Blockchain for Future Cybersecurity Applications*.:217.
- [3] Siam AI, Almaiah MA, Al-Zahrani A, Elazm AA, El Banby GM, El-Shafai W, El-Samie FE, El-Bahnasawy NA. Secure Health Monitoring Communication Systems Based on IoT and Cloud Computing for Medical Emergency Applications. *Computational Intelligence and Neuroscience*. 2021 Dec 13;2021.
- [4] Al Nafea R, Almaiah MA. Cyber security threats in cloud: literature review. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 779-786). IEEE.
- [5] AlMedires M, AlMaiah M. Cybersecurity in Industrial Control System (ICS). In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 640-647). IEEE.
- [6] Alamer M, Almaiah MA. Cybersecurity in Smart City: A systematic mapping study. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 719-724). IEEE.
- [7] Ali A, Almaiah MA, Hajje F, Pasha MF, Fang OH, Khan R, Teo J, Zakarya M. An Industrial IoT-Based Blockchain-Enabled Secure Searchable Encryption Approach for Healthcare Systems Using Neural Network. *Sensors*. 2022 Jan;22(2):572.
- [8] Almudaires F, Almaiah M. Data an overview of cybersecurity threats on credit card companies and credit card risk mitigation. In2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 732-738). IEEE.
- [9] Almaiah A, Almomani O. An investigation of digital forensics for shmoon attack behaviour in FOG computing and threat intelligence for incident response. *Journal of Theoretical and Applied Information Technology*. 2020 Apr 15;98(07).
- [10] Qasem MH, Obeid N, Hudaib A, Almaiah MA, Al-Zahrani A, Al-Khasawneh A. Multi-Agent System Combined With Distributed Data

- Mining for Mutual Collaboration Classification. *IEEE Access*. 2021 Apr 20;9:70531-47.
- [11] ALMAIAH A, Almomani O. An Investigator Digital Forensics Frequencies Particle Swarm Optimization for Detection And Classification of Apt Attack in Fog Computing Environment (IDF-FPSO). *Journal of Theoretical and Applied Information Technology*. 2020 Apr 15;98(07).
- [12] Almaiah MA. An Efficient Smart Weighted and Neighborhood-enabled Load Balancing Scheme for Constraint Oriented Networks.
- [13] Almaiah MA, Al-Zahrani M. Multilayer Neural Network based on MIMO and Channel Estimation for Impulsive Noise Environment in Mobile Wireless Networks. *International Journal of Advanced Trends in Computer Science and Engineering*. 2020;9(1):315-21.
- [14] Adil M, Khan R, Almaiah MA, Al-Zahrani M, Zakarya M, Amjad MS, Ahmed R. MAC-AODV based mutual authentication scheme for constraint oriented networks. *IEEE Access*. 2020 Mar 4;8:44459-69.
- [15] Adil M, Khan R, Ali J, Roh BH, Ta QT, Almaiah MA. An energy proficient load balancing routing scheme for wireless sensor networks to maximize their lifespan in an operational environment. *IEEE Access*. 2020 Aug 31;8:163209-24.
- [16] Bubukayr MA, Almaiah MA. Cybersecurity concerns in smart-phones and applications: A survey. In 2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 725-731). IEEE.
- [17] Almomani O, Almaiah MA, Alsaaidah A, Smadi S, Mohammad AH, Althunibat A. Machine Learning Classifiers for Network Intrusion Detection System: Comparative Study. In 2021 International Conference on Information Technology (ICIT) 2021 Jul 14 (pp. 440-445). IEEE.
- [18] Al Hwaitat AK, Almaiah MA, Almomani O, Al-Zahrani M, Al-Sayed RM, Asaifi RM, Adhim KK, Althunibat A, Alsaaidah A. Improved security particle swarm optimization (PSO) algorithm to detect radio jamming attacks in mobile networks. *Quintana*. 2020;11(4):614-24.
- [19] Almaiah MA, Dawahdeh Z, Almomani O, Alsaaidah A, Al-khasawneh A, Khawatreh S. A new hybrid text encryption approach over mobile ad hoc network. *International Journal of Electrical and Computer Engineering (IJECE)*. 2020 Dec;10(6):6461-71.
- [20] Adil M, Khan R, Almaiah MA, Binsawad M, Ali J, Al Saaidah A, Ta QT. An efficient load balancing scheme of energy gauge nodes to maximize the lifespan of constraint oriented networks. *IEEE Access*. 2020 Aug 11;8:148510-27.
- [21] Khan MN, Rahman HU, Almaiah MA, Khan MZ, Khan A, Raza M, Al-Zahrani M, Almomani O, Khan R. Improving energy efficiency with content-based adaptive and dynamic scheduling in wireless sensor networks. *IEEE Access*. 2020 Sep 25;8:176495-520.
- [22] Adil M, Almaiah MA, Omar Alsayed A, Almomani O. An anonymous channel categorization scheme of edge nodes to detect jamming attacks in wireless sensor networks. *Sensors*. 2020 Jan;20(8):2311.
- [23] D. P. F. Möller and R. E. Haas, *Automotive Cybersecurity: A Guide to Automotive Connectivity and Cybersecurity*. Cham: Springer International Publishing, pp. 265–377, 2019.
- [24] M. Waheed and M. Cheng, "A system for real-time monitoring of cybersecurity events on aircraft," *IEEE*, pp. 1–3, 2017.
- [25] E. Aboul, M. Hassanién and Elhoseny, *Advanced Sciences and Technologies for Security Applications Cybersecurity and Secure Information Systems Challenges and Solutions in Smart Environments*.
- [26] A. Zarreh, H. Wan, Y. Lee, C. Saygin and R. A. Janahi, "Cybersecurity concerns for total productive maintenance in smart manufacturing systems," *Procedia Manufacturing*, vol. 38, pp. 532–539, 2019.
- [27] G. Pizzi, "Cybersecurity and its integration with safety for transport systems: Not a formal fulfillment but an actual commitment," *Transportation Research Procedia*, vol. 45, pp. 250–257, 2020.
- [28] M. Kevin, F. Ana and K. Alexey, "Smart information systems in cybersecurity," *The ORBIT Journal*, vol. 2, no. 2, pp. 1–26, 2019.
- [29] C. Hess. and E. Ostrom, "A Framework for Analyzing the Knowledge Commons: a chapter from Understanding Knowledge as a Commons: from Theory to Practice, 2005," *Conference on Dependability and Complex Systems 2015*, pp. 97–106.
- [30] K. Cheung, M. G. Bell and J. Bhattacharjya, "Cybersecurity in logistics and supply chain management: An overview and future research directions," *Transportation Research Part E: Logistics and Transportation Review*, vol. 146, p. 102217, 2021.
- [31] P. Lau, W. Wei, L. Wang, Z. Liu and C. Ten, "A cybersecurity insurance model for power system reliability considering optimal defense resource allocation," *IEEE Transactions on Smart Grid*, vol. 11, no. 5, pp. 4403–4414, 2020.
- [32] A. Thaduri, M. Aljumaili, R. Kour and R. Karim, "Cybersecurity for e-maintenance in railway infrastructure: Risks and consequences," *International Journal of System Assurance Engineering and Management*, vol. 10, no. 2, pp. 149–159, 2019.
- [33] E. Kweon, H. Lee, S. Chai and K. Yoo, "The utility of information security training and education on cybersecurity incidents: Empirical evidence," *Information Systems Frontiers*, pp. 1–13, 2019.
- [34] R. Baskerville, F. Rowe and F. Wolff, "Integration of information systems and cybersecurity countermeasures: An exposure to risk perspective." *ACM SIGMIS Database: The Database for Advances in Information Systems*, vol. 49, no. 1, pp. 33–52, 2018.
- [35] Q. Zhang, C. Zhou, N. Xiong, Y. Qin, X. Li and S. Huang, "Multimodel-based incident prediction and risk assessment in dynamic cybersecurity protection for industrial control systems," *IEEE Transactions on Systems, Man, and Cybernetics. Systems*, vol. 46, no. 10, pp. 1429–1444, 2016.
- [36] N. Polatidis, M. Pavlidis and H. Mouratidis, "Cyberattack path discovery in a dynamic supply chain maritime risk management system," *Computer Standards & Interfaces*, vol. 56, pp. 74–82, 2018.
- [37] M. Alali, A. Almogren, M. M. Hassan, I. A. L. Rassa and M. Z. A. Bhuiyan, "Improving risk assessment model of cyber security using fuzzy logic inference system," *Computers & Security*, vol. 74, pp. 323–339, 2018.
- [38] J. Van Erp, "New governance of corporate cybersecurity: A case study of the petrochemical industry in the Port of Rotterdam," *Crime, Law and Social Change*, vol. 68, no. 1, pp. 75–93, 2017.
- [39] G. Lykou, G. Iakovakis and D. Gritzalis, "Aviation cybersecurity and cyberresilience: assessing risk in air traffic management," *Critical Infrastructure Security and Resilience*. Cham: Springer International Publishing, pp. 245–260, 2019.
- [40] I. De La Peña Zarzuelo, "Cybersecurity in ports and maritime industry: Reasons for raising awareness on this issue," *Transport Policy*, vol. 100, pp. 1–4, 2021.

# ASM-ROBOT: A Cyber-Physical Home Automation Controller with Memristive Reconfigurable State Machine

Kennedy Chinedu Okafor<sup>1</sup>, Omowunmi Mary Longe<sup>2</sup>

Dept. of Mechatronics Engineering, Federal University of Technology, Owerri, Nigeria<sup>1</sup>  
Electrical and Electronic Engineering Science, University of Johannesburg, South Africa<sup>1,2</sup>

**Abstract**—In the next 5 to 10 years, digital Artificial Intelligence with Machine Circuit Learning Algorithms (MCLA) will become the mainstream in complex automated robots. Its power concerns, ethical perspectives, including the issues of digital sensing, actuation, mobility, efficient process-computation, and wireless communication will require advanced neuromorphic process variable controls. Existing home automated robots lack memristic associative memory. This work presents Cyber-Physical Home Automation System (CPHAS) using Memristive Reconfigurable Algorithmic State Machine (MRASM) chart. A process control architecture that supports Concurrent Wireless Data Streams and Power-Transfer (CWDSPT) is developed. Unlike legacy systems with power-splitting (PS) and time-switching (TS) controls, the MRASM-ROBOT explores granular wireless signal controls through unmodulated high-power continuous wave (CW). This transmits infinite process variables using Orthogonal Space-Time Block Code (OSTBC) for interference reduction. The CWDSPT transmitter and receiver circuit design for signal processing are implemented with complexity noise-error reduction during telemetry data decoding. Received signals are error-buffered while gathering control variables' status. Transceiver Memristive neuromorphic circuits are introduced for computational acceleration in the design. Hardware circuit design is tested for system reliability considering the derived schematic models for all process variables. Under small range space diversity, the system demonstrated significant memory stabilization at the synchronous iteration of the synaptic circuitry.

**Keywords**—Cloud computing; cyber-physical systems; complex robot; computational science; IoT; machine learning

## I. INTRODUCTION

### A. Background

Memristive Reconfigurable Algorithmic State Machine Robot (MRASM-ROBOT) is a novel intelligent automation approach that provides support for the control of converged appliances using localized space diversity. The concept is very useful for providing security to both localized and remote homeowners. It exploits associative memory which has a structured memory unit for storing identified data both at the edge and cloud control levels respectively. The system uses access by data content (content addressable memory), and address by memory location (associative memory).

Memristive neuromorphic circuits can be constructed with associative memory for infinite process variables such as

temperature, smoke, flames, gases, humidity, and water, among others. Modern-day hardware modules can easily be implemented using Memristive reconfigurable memory systems. This is found in neuromorphic robots and other complex systems [1]. The application areas are very huge, thereby creating demands for location-independent automation processes. Smart Internet of Things (IoT) is often used in this regard due to space diversity considerations. In such cases, IoT RF technology (such as ATMEL Smart-SAM25 module) with the CWDSPT technique can be adapted to provide effective monitoring and control of both on-premises and remote processes. With heightened security challenges as well as domestic havoc at homes and offices, such an automated security system becomes very handy [1], [2].

Existing home automation approaches fail to account for space diversity sensitivity. Also, process variable coordination in various application contexts is unpredictable. In most legacy systems, their efficiency often involves a high cost, especially when using remote network services. There is a need for an automated security system that is optimized for high efficiency in controlling home appliances at short-range communication. Smart Energy Audit Systems (SEACS) have been proposed which largely depends on IoTs via near field communication (NFC) [2]. Hand-gesture-based control Robots have been implemented for real-time interactive control systems targeted at household appliances [3]. Home automation systems for anomaly classification using unsupervised probabilistic associations for sensing and event tracing in smart homes are trending [4], [5], and [6].

Recently, researchers are making efforts to drive biological neural networks and map with memristive neuromorphic models to customize AI-based Robots [7], [8], [9]. Similarly, memristive neuromorphic models have been applied in most computing acceleration designs [10], [11], [12]. In classical literature such as social psychology, the capacity to learn and recall the relationship-map existing among unconnected variables is referred to as associative memory [13]. This could be declarative in its structure or episodic [14] in its application contexts.

Various attributes of complex radio frequency (RF) robots such as real-time withdrawal reflex, classical eyeblink, etc., can occur via a complex learning process in associative memory. In this case, a conditioned reflex can be established between conditioned stimulus (CS) and system state-response.

Low range RF memristive neuromorphic circuits can be leveraged as a Pavlov associative memory (PAM) or a withdrawal reflex [15], [16]. Max-input-feedback adaptive-learning rule has been applied to train the neuromorphic circuit while adapting it to PAM [17]. Some designs explore microcontrollers, and signal conditioners such as A/D converter to synthesize memristor features in a robot. In this case, the memristor can be used as a synaptic/ neuromorphic circuit, to model PAM [18]. Though finite state machine is not new, however, it can be employed to assist an unsupervised  $k$ -means logic intelligence while monitoring and detecting abnormal conditions.

### B. Advantages of Memristive Control Design Technique

- Offers bifurcation analysis involving two or more parameters and predict the linear stability boundaries.
- Similar to biological synapses, dendrites, and neurons, the memristive technique with neuromorphic computing offers another layer to AI at a physical level.
- The memristive method offers insights into non-linear systems by extending the capabilities of resistors, Capacitors, and Inductors.
- The Memristive approach provides optimal control in circuit diversity.

Existing works on memristive systems focused on high-density filters or FPGAs volatile memory, crossbar latches, neural networks, modeling of neural synapses, nonlinear oscillators, and filters. Very little work has been done field of Memristive reconfigurable state machine and space diversity control signaling.

### C. Research Contributions

This work presents a novel control strategy for process variable transceiver modules in smart homes designs. The system allows for the optimization of errorless information decoding. The other contributions in this research include:

- Derive a standardized remote-controlled automation architecture adapted for process aggregation dynamics.
- To evaluate the IoT-gateway transmission and receiver behavior for a low bit error rate under the influence of Multipath Rayleigh Channel Space Vector (MRCSV) and White Gaussian noise (AWGN).
- To show the impact of space diversity on received signal using optimal OSTBC-combiner block (CB).
- To demonstrate MRASM-ROBOT use case scenario for bit error response at scale.

The rest of the paper is organized as follows. Section II presents related works. Section III presents the methodology. Section IV presents the smart automation engine (SAE). Section V presents the feedback scheduling algorithm, Section VI focused on space diversity control. Section VII discussed the experimental results, while Section VIII concludes the work.

## II. RELATED WORK

Mostly, there are challenges regarding the coordination of control states and the collection of telemetry data for smart home analytics. The absence of MRASM-CWDSPT affects the accuracy of data communication on the home server. This leads to under-utilization and lower efficiency. In this section, closely related literatures on two important elements were studied namely: CWDSPT Space diversity and automated/smart homes without neuromorphic circuits.

### A. CWDSPT Space Diversity

Concurrent Wireless Data Streams and Power-Transfer is an emerging area with great prospects especially with 5G rollout in Africa. In most works, spatial diversity (SD) is introduced in transmissions systems to mitigate the effect of multipath fading [19]. In lengthy links over extremely reflective surfaces like water bodies where a non-diversity link can't deliver high availability, SD is needed. It is used in multiple antennas that have similar characteristics involving real physical separation from each other. There have been several works on space diversity, but the work in [20] used a concave-convex procedure (CCCP) scheme to maximize the minimum rate of IoT nodes' private streams. This was done via the allocation of transmit power and adjustment of power-splitting ratios at the IoT edge nodes.

A similar work [21], discussed spatially modulated space-time block code (HRSM-STBC) developed for two active antennas. To optimize space diversity schemes, extensive discussions on wireless power transfer (WPT) capabilities were presented in [22]. The work [23] investigated a wireless power transfer (DWPT) tool for developing wireless power transfer (WPT) systems and its related parameters. In [24], the authors proposed a radial-flux rotational wireless power transfer (RF-R-WPT) system with a rotor state identification function for devices mounted on a rotating shaft. In [25], the authors focused on the systematic review of metamaterials and meta surfaces for wireless power transfer (WPT) and wireless energy harvesting (WEH) [26].

In [27], the authors leveraged Wireless power transfer (WPT) systems due to their security, convenience, and flexibility to propose a control strategy referred to as periodic energy control (PEC) in WPT. So far, current literature on wireless information power transfer and space diversity has not been explored in Cyber-physical home automation. Its applications in MRASM-ROBOT will be very novel and useful too.

### B. Neuromorphic State Machine (NSM)

The authors [28], developed an inverter-based memristive neuromorphic RF switch on the microstrip line with good performance up to 5GHz. Their design explored conductive bridging random access memory (CBRAM) scheme in which the switch uses metal-insulator-metal (MIM) structure, Copper/Nafion/Au majorly. In [29], the authors focused on inverter-based memristor crossbar neural networks. The work [30] investigated the impacts associated with the computing accuracy of analog memristive circuits for neuromorphic applications. In [31], the security application of memristive crossbar physical unclonable function was investigated for

resource optimization. The work [32] investigated neuromorphic computational system models using the spine technique. This is used to improve system lifetime in mapping machine learning workloads. The work in [33] focused on the optimization of Memristive Crossbar Arrays. In [34], the authors investigated memristance variations capture synaptic weight variations. The authors [35], [36] proposed the Pavlov associative memory process otherwise known as multi-functional memristive Pavlov associative memory circuit. The work [37] proposed a weight optimization scheme that combines quantization and Bayesian inference for memristor-based neuromorphic computing system (NCS).

In cases of Mixed-signal analog/digital neuromorphic circuits, there has been the characterization of such circuits using ultra-low power consumption, real-time processing abilities, and low-latency response times [38]. In this case, neuromorphic processors are used as the neuromorphic agent. The work [39] discussed a new neuro-inspired, hardware-friendly readout stage for the liquid state machine (LSM) using neuromorphic VLSI implementation. The work [40] discussed a neuromorphic computational scheme known as the prefrontal cortex (PFC). The work is equally looked at as mixed-signal analog/digital neuromorphic implementation. Authors [41] highlighted that memristive circuits can trigger the associative memory processes while retaining the design parameters.

There are other types of memristive circuits without any clear synaptic and neuron circuits. Hence, with a modified memristive circuit, the memristive neuromorphic models can be employed to enhance computational efficiency in deployment contexts.

Considering the memristive neuromorphic circuits used for computation acceleration, the work [42], [43], [44], showed that the synaptic circuit is required to represent process variable weights in binary form. The works equally highlighted that various synaptic circuits can be designed to utilize neuromorphic circuits. At the implementation level, Field programmable gate arrays (FPGAs) have been used to provide neuromorphic reconfigurable characteristics while enabling flexible schematic designs for various applications [44], [45], [46], [47]. Interestingly, various control circuit topologies designated for process variables in existing designs lack the functionality of MRASM states. Also, the design of memristive neuromorphic circuits must replicate the process of learning, with high retention for conditional state-based signals.

Therefore, the gap in NSM is that low-end cyber-physical applications cannot fit into the design complexity due the memory stabilization issues. However, the use of associative memory can help in the design of Cyberphysical robots that learns and remembers the relationship between unrelated things. Such reconfigurable NSM can be useful in the design of intelligent RF systems.

### C. Summary of Related Works (Automated Home)

The use of NSM to drive space diversity intelligent systems is novel. In this section, the summary current related efforts on automated/smart home without NSM. In [48], the authors proposed Open-source home automation systems which lacked a detailed implementation framework. In the work [49], the

authors proposed home energy management systems (HEMS). The design supports end-users by allowing for demand-side management and automation. In [50], the authors developed a secured but lightweight three-factor driven privacy-preserving authentication model specifically for IoT-enabled smart home environments. The authors [51], presented a scheme that automates home appliances in three modes viz: local, web, and app-based automation. These were achieved with a low-cost microcontroller, a web page with support for remote applications. The work [52] presented an improved robots index model used to optimize the robust level of home energy local network (HELN) while considering its numerous household appliances. In [53], a novel proximity service model for smart home automation is proposed. Their work uses wireless networks and native Internet connectivity. The work [54] discussed an asynchronous electrooculography (EOG)-based human-machine interface (HMI) for smart home environmental control. The work [55] focused on a machine-learning-based approach to assessing activity quality in smart homes based on automation assessment. The work [56] highlighted an efficient implementation approach using IoT for real-time monitoring of routine activities in homes.

So far, the future of ubiquitous home networks will rely on sensors to aggregate various environmental data variables [57]. This work has similar baseline design attributes with robotic smart homes based on stabilized feedback Episodic memory (SF-EM) [58], stewards robot smart homes [59], humanoid defense smart homes [60], Indoor autonomous robots [61], and Smart home activities IoT [62]. These similarities are highlighted below [63], [64], [65]:

- Process automation through task allocation.
- Lightweight process variable control.
- Layered integrations with NSM.

The identified gaps/limitations are found below:

- Absence of space diversity characterization under multipath fading channels.
- Absence of NSM for hardware analytics.
- Though most systems have complex design topologies.
- Absence of CPS process variable control using the associative-memory based on the reconfigurable memristive neuromorphic scheme.
- The absence of multipath fading channel optimization for memory stabilization is novel.

### D. Summary of Contributions

The generalized smart home systems are anticipated to provide custom-based services, especially to home users thereby reducing human efforts. Unfortunately, the legacy computational knowledge-driven algorithms for learning and reasoning do not completely orchestrate unpredictable changes in these homes. In this work, the service provisioning demand seeks to handle: i.) remote learning and reasoning process control algorithms, and ii) Edge layer integration of smart endpoints. To fix these issues, the contributions of this work are as follows.

- 1) Develop an associative-memory-induced reconfigurable memristive neuromorphic for memory stabilization.
- 2) Introduce a logical feedback mechanism for behavioral learning from the state machine.
- 3) Complete a set of space diversity optimization for memory service guarantee from multipath fading errors.

### III. METHODOLOGY

The physical design topology of the smart automation engine (SAE) is shown in Fig. 1. This depicts the design symmetry of MRASM-ROBOT enabled with machine learning automation and CWDSPT. The implementation of the smart automation/security system is based component layered approach and its subsystems were implemented as a stand-alone system for ease of reconfiguration. The optimally controlled process variables modeled are temperature states, room lighting, overhead tank, as well as house security.

The design can protect the occupants from life-threatening hazards. This is achieved by getting inputs from sensors placed at different locations throughout the house/deployment environment. Application program interfaces (API) are used for remote communications.

The inputs are then fed into the Arduino MKR1000 Wi-Fi process controller (PCon) through which these different aspects of the house are controlled. Essentially, communication happens in a short-range transmission distance of 100 to 100mW. This class 1 range uses the RF ISM bands (2.402GHz-2.48GHz).

The objective function (OF) is the infinite process maximization with constraint variables such as link capacity, channel effects, resource availability, and cost function. These are responsible for the smart automation engine (SAE). This further incorporates static real feedback looping for dynamic stability, having an optimal algorithm with low-bit error signaling. These functionalities are implemented to monitor and control devices that drain energy while securely protecting the home facility. As shown in Fig. 1, there are two major parts viz.

- Automation module. This controls the lights, inductive load, and cooling systems (HVACs) in and around the house thereby saving energy. Also, it has the following sub-systems namely – the process controller, level converter, GSM modem, smart green power supply unit, the input interface (sensor subcircuit), control system, and the output Interface such as APIs.
- Associative-memory-induced scheme i.e., memristive reconfigurable state machine attributes.

A brief explanation of the resilient transmitter, RF-wireless communication channel (i.e., transmitter/receiver) for the proposed MRASM-CWDSPT is presented via the component modeling technique. Also, the details proposed CWDSPT, (i.e., DC-biased orthogonal frequency-division multiplexing (DC-biased OFDM) in the IoT controller is equally presented below.

Let's define the following process variables as follows:

$API_{g_m} = API\ gateway$

$SGg_{v_\alpha} = smaoke\ signal\ conditioner\ API\ communication$

$WT_{Cm_\beta} = water\ signal\ conditioner\ API\ communication$

$TM_{Cm_\beta} = Temperature\ signal\ controller\ API\ communication$

$LG_{as_\mu} = Light\ intensity\ siganl\ controller\ API\ communication$

The sink controller functions include Smoke, water, temperature, and light intensity.

#### A. Memristive Reconfigurable Neuromorphic Circuit Model

In this section, the component modeling technique was introduced to build the MRASM-Robotic subsystem. This involves process variable synaptic modeling. First, a neuron circuit refers to the schematic baseline unit used to implement a functional neuromorphic system for MRASM-Robotic. It connects more than one synaptic subcircuit interface module depicted in Fig. 1.

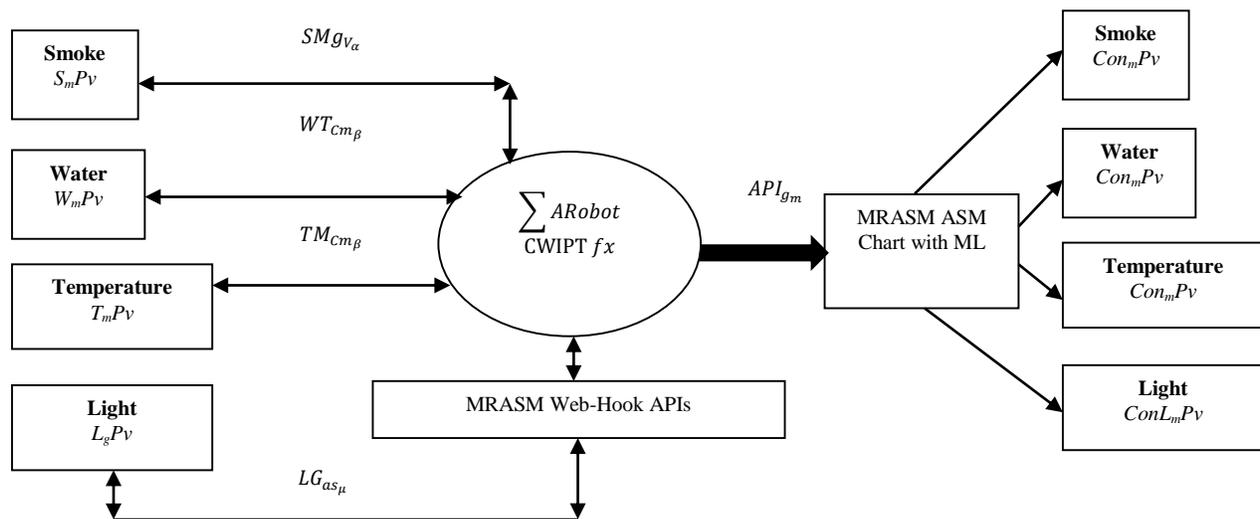


Fig. 1. MRASM-ROBOT Process Model for Memory Stabilization.

Second, there are three major parts in the neuron circuit: 1) MRASM-Robotic selection circuitry; 2) MRASM-Robotic synaptic circuitry, and 3) MRASM-Robotic neuron circuitry.

In the design implementation, the MRASM-Robotic neuron circuitry has two input voltage levels/signals, MVanalog and MVdigital. The former depicts the analog signal estimated at 1–5 V. The latter refers to the binary digital signal where the voltages of logic 1 and logic 0 are represented as 5 and 0 V respectively. Voltage thresholds of pMOS transistors and nMOS transistors in MRASM-Robot are fixed at  $-0.6$  V and  $0.6$  V, respectively.  $T_1$  and  $T_2$  pMOS are used for the selection circuitry.

Now, let's define the computational core of the MRASM-Robot which orchestrates both the input and output subsystems in Fig. 1. The process variables (smoke, water, temperature, light signals) modeled in the design include the overhead tank water level sensor module ( $W_mPv$  in Fig. 2a), a Smoke detection module ( $S_mPv$  in Fig. 2b), temperature control system ( $T_mPv$  in Fig. 2c), and motion-controlled lighting ( $L_gPv$  in Fig. 2d.).

Using digital comparators shown in Fig. 2a-d, the work estimated appropriate voltages at the output of the comparator. The water resistance was determined experimentally by using a neural network digital ohmmeter (NNDM). A resistance of approximately between  $100\text{ k}\Omega$  and  $120\text{ k}\Omega$  was obtained using various samples of water. The resistance of  $100\text{ k}\Omega$  was taken to be the water resistance using voltage divider models while accounting for tolerance consideration.

A tamper-proof was introduced to monitor open circuit situations or when the sensor goes bad. This is included since if the sensor goes bad without being detected, lives and properties may be lost. The system consists of an additional length of wire connected along with the smoke sensor. This drives the transistor to saturation and diode  $D_1$  then conducts. When there is a tamper or if the sensor goes bad, this condition is sensed by the processing unit. From Fig. 2c, ideally, the input of an Op-amp has almost infinite resistance. This then forces the  $V_{cc}$  to be dropped across  $R_{30}$  and  $R_{31}$ . Since the output of LM35 is between  $0\text{ V}$  to  $1\text{ V}$  ( $10\text{ mV}/^\circ\text{C}$  to be precise). To allow power drop across  $R_{31}$ , the value was arbitrarily chosen as  $1\text{ M}\Omega$ . This is well shown in Fig. 2c.

Furthermore, the use of empirical data from machine learning computations is currently investigated in Fig. 1. This is to show the extent of metrics performance between the Memristic control strategy and the traditional models using quantified datasets. Also, The adjusting/changing memristor states are done by changing the resistance value over time during switching of control state variables.

### B. Motion Controlled Lighting

For the LDR section, various experimental trials were carried out. It was noticed that in a dark environment, the LDR will have a resistance of  $10\text{ k}\Omega$ , thus allowing  $2.5\text{ V}$  to drop across LDR as shown in Fig. 2d. That means that the voltage drop across  $R_{49} = 2.5\text{ V}$ . The essence of this sensing arrangement is to create a path across which blockage will obstruct the transmitted rays and the receiving phototransistor, resulting in an open circuit. Anybody entering or leaving the

room as a result of crossing this path creates such a situation and will be detected.

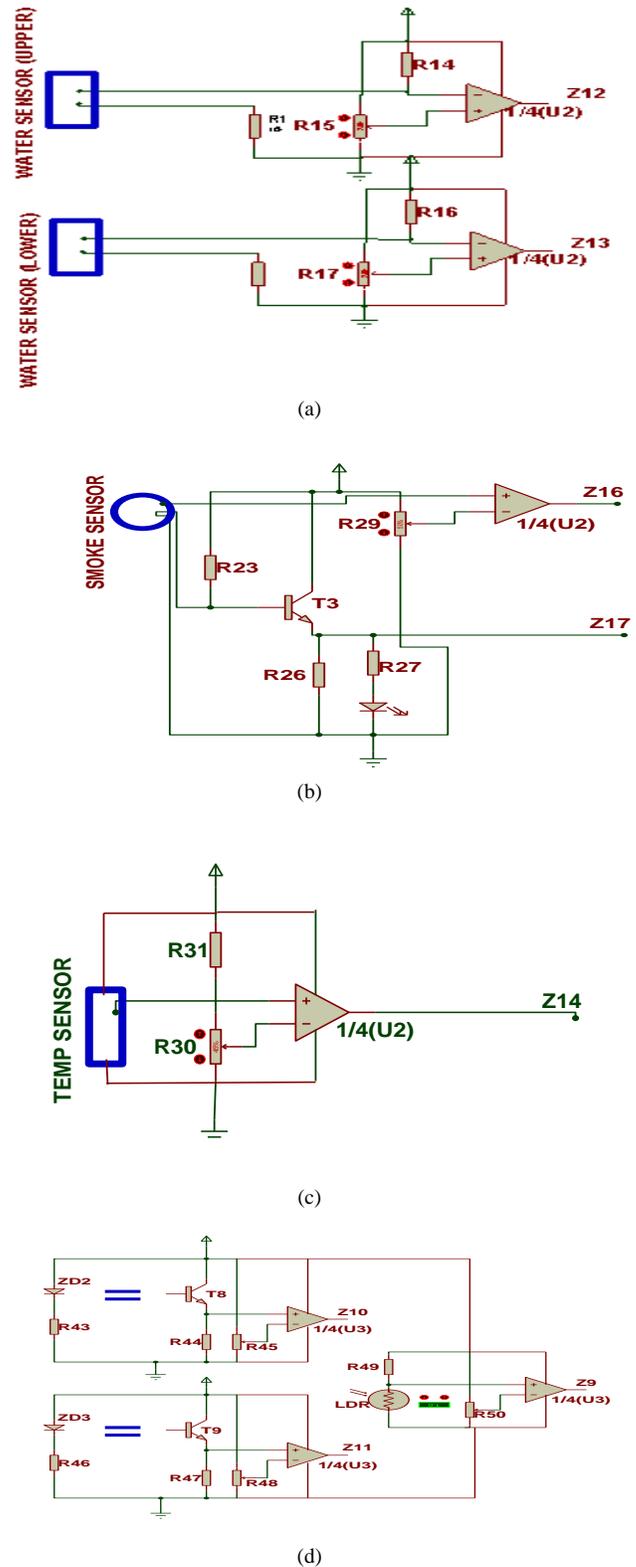


Fig. 2. (a) MRASM Overhead Tank Sensor, (b). MRASM Smoke Detection Module, (c): MRASM Temperature Sensor, (d). MRASM Motion-controlled Light.

### C. Pass-worded Locking System

Push switches are used like keypads for security controls. A switch is used with a resistor as shown in Fig. 3. The value of the resistor selected is 10 kΩ resistor. When the switch is 'open' the 10 kΩ resistor connects the microcontroller input pin down to 0 V, which gives an off (logic level 0) 0 V signal to the microcontroller input pin. When the switch is activated, the input pin is connected to the positive battery supply (V+). This provides an on (logic level 1) signal to the PCon. It is a 4x4 matrix keypad requiring eight Input/Output ports for interfacing. Rows are connected to peripheral Input/Output (PIO) pins configured as an output. Columns are connected to PIO pins configured as input with interrupts. In this configuration, four pull-up resistors must be added to apply a high level on the corresponding input pins. The corresponding hexadecimal value of the pressed key is sent on four LEDs.

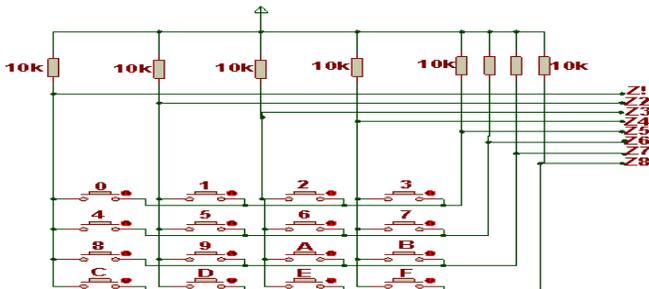


Fig. 3. MRASM-ROBOT Keypad Interface.

## IV. SMART AUTOMATION ENGINE

In this Section, MRASM-ROBOT was improved with feedback scheduling at the controller level. PIDs algorithm was added for dynamic stability and process controls. The IoT space diversity model for reliable data stream/message delivery is developed. This uses a two-way transceiver communication for memory stabilization. Static real-time scheduling for the automation processes in MRASM-ROBOT was introduced. Also, the design optimization algorithm and implementation emulations, among others were achieved via proofs-of-concept. These components comprise the smart automation engine for MRASM-ROBOT.

### A. Design Description

- Considering the block diagram of the smart automation engine (SAE) in Fig. 1, three layers were identified viz. core, access/automation layer, and aggregation layers. Recall that the process controller is the major component in the core speed redundancy layer, while automation devices in the access layer are interconnected through the ports of the process controller. The MRASM-Robot automation shown in Fig. 4 has the model specifications briefly discussed. There are six process variables in the access layer. These include Water Level Signal ( $P_{v1\_WLS}$ ), Temperature Control Signal ( $P_{v2\_TCS}$ ), Smoke Signal ( $P_{v3\_SKS}$ ), Motion Detection Signal ( $P_{v4\_MDS}$ ), = Keypad Signal ( $P_{v5\_KPS}$ ). These are all interconnected as in Fig 4. In programming script,  $P_{v1}$ ,  $P_{v2}$ ,  $P_{v3}$ ,  $P_{v4}$ ,  $P_{v5}$  denotes Water Level Signal (WLS), Temperature Control Signal (TCS), Smoke Signal (SKS), Motion

Detection Signal (MDS), and Keypad Signal (KPS) among others. These are all connected to the Process Controller (PCon).

- The process controller uses the optimization algorithm in the aggregation layer to recursively monitor and control the devices without a drop in their stability.
- The MRASM-ROBOT output interface is modeled to report recurrent processes on the system display dashboard or sink. Commands instruction is given to dashboard then execute any predefined task such as initialization, reset, setting the cursor position, controlling the display, etc. The data register stores the data to be displayed on the dashboard. The data is the ASCII value of the character to be displayed on the dashboard.

### B. Smart Automation MRASM Chart

In the MRAM automation system, the derived chart was designed which controls the process variables shown in Fig. 4 using wireless signaling (i.e. CWDSPT). The MRASM chart design is essentially is a finite state machine design scheme used to represent diagrams of process logic regulation in the PCon. The MRASM technique comprises these steps:

- 1) Create pseudocode for the desired operation of MRASM on the controller.
- 2) Translate the pseudocode into an MRASM chart.
- 3) Derive the datapath from the MRASM chart.
- 4) Create a detailed MRASM chart based on the Datapath.
- 5) Create the control logic from the MRASM chart.

The outcome of the above steps resulted in Fig. 5 showing the PCon MRASM logic box for the smart automation system. Fig. 5 shows the PCon MRASM logic box for the smart automation system. Fig. 6 shows the design flowchart.

For the model, let the process variables be represented in the MRASM chart State Transition Table presented in Table I with various variables defined below:

$P_{v1\_WLS}$  = Water Level Signal;  $P_{v2\_TCS}$  = Temperature Control Signal;  $P_{v3\_SKS}$  = Smoke Signal;  $P_{v4\_MDS}$  = Motion Detection Signal;  $P_{v5\_KPS}$  = Keypad Signal

PSN = Present State Name; PSC = Present State Code; NSN = Next State Name; NSC = Next State Code.

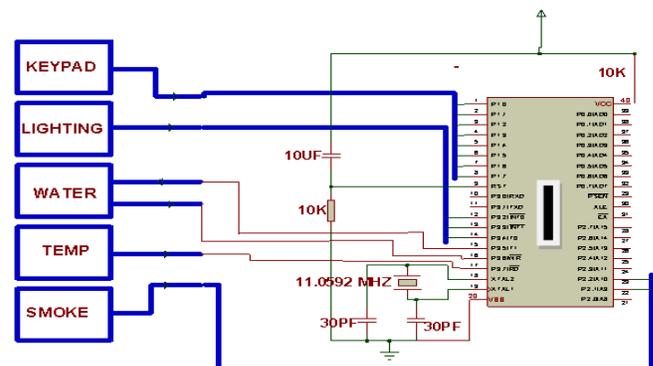


Fig. 4. MRASM-ROBOT Input Interfaces for  $P_{v1}$ ,  $P_{v2}$ ,  $P_{v3}$ ,  $P_{v4}$ ,  $P_{v5}$ , &  $P_{vn}$ .

TABLE I. MEMRISTIVE RECONFIGURABLE STATE MACHINE TRANSITION

| LINK PATH | Input Qualifiers |     |     |     |     | PSN DCBA | PSC  | NSN DCBA | NSC  | Output |     |     |     |     |
|-----------|------------------|-----|-----|-----|-----|----------|------|----------|------|--------|-----|-----|-----|-----|
|           | WLS              | TCS | SKS | MDS | KPS |          |      |          |      | RL1    | RL2 | BUZ | RL3 | RL4 |
| L1        | 0                | -   | -   | -   | -   | ST0      | 0000 | ST0      | 0000 |        |     |     |     |     |
| L2        | 1                | -   | -   | -   | -   | ST0      | 0000 | ST1      | 0001 | 1      |     |     |     |     |
| L3        | 1                | -   | -   | -   | -   | ST1      | 0001 | ST1      | 0001 |        |     |     |     |     |
| L4        | 0                | -   | -   | -   | -   | ST1      | 0001 | ST2      | 0011 | 0      |     |     |     |     |
| L5        | -                | 0   | -   | -   | -   | ST2      | 0011 | ST2      | 0011 |        |     |     |     |     |
| L6        | -                | 1   | -   | -   | -   | ST2      | 0011 | ST3      | 0010 |        | 1   |     |     |     |
| L7        | -                | 1   | -   | -   | -   | ST3      | 0010 | ST3      | 0010 |        |     |     |     |     |
| L8        | -                | 0   | -   | -   | -   | ST3      | 0010 | ST4      | 0110 |        | 0   |     |     |     |
| L9        | -                | -   | 0   | -   | -   | ST4      | 0110 | ST4      | 0110 |        |     |     |     |     |
| L10       | -                | -   | 1   | -   | -   | ST4      | 0110 | ST5      | 0111 |        |     | 1   |     |     |
| L11       | -                | -   | 1   | -   | -   | ST5      | 0111 | ST5      | 0111 |        |     |     |     |     |
| L12       | -                | -   | 0   | -   | -   | ST5      | 0111 | ST6      | 0101 |        |     | 0   |     |     |
| L13       | -                | -   | -   | 0   | -   | ST6      | 0101 | ST6      | 0101 |        |     |     |     |     |
| L14       | -                | -   | -   | 1   | -   | ST6      | 0101 | ST7      | 0100 |        |     |     | 1   |     |
| L15       | -                | -   | -   | 1   | -   | ST7      | 0100 | ST7      | 0100 |        |     |     |     |     |
| L16       | -                | -   | -   | 0   | -   | ST7      | 0100 | ST8      | 1100 |        |     |     | 0   |     |
| L17       | -                | -   | -   | -   | 0   | ST8      | 1100 | ST8      | 1100 |        |     |     |     |     |
| L18       | -                | -   | -   | -   | 1   | ST8      | 1100 | ST9      | 1101 |        |     |     |     | 1   |
| L19       | -                | -   | -   | -   | 1   | ST9      | 1101 | ST9      | 1101 |        |     |     |     |     |
| L20       | -                | -   | -   | -   | 0   | ST9      | 1101 | ST0      | 0000 |        |     |     |     | 0   |

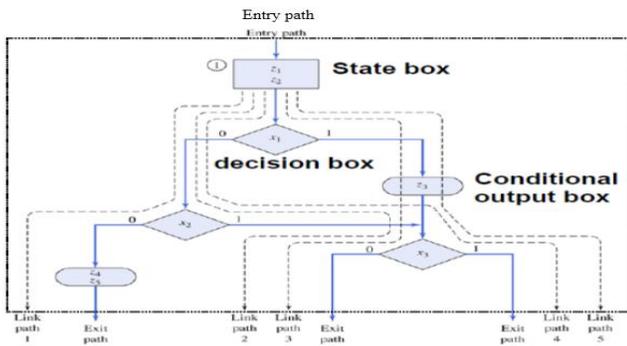


Fig. 5. MRASM Process Controller Block.

C. MRASM Design Description

As depicted in Table I, the design chart for MRASM smart automation security control was further characterized using schematics capture C++ scripting. Test case values are derived from optimal system response conditions. The input interface to the PCon is made up of the input qualifiers. The qualifiers are the comparator output from the different sensor modules at the access layer. As shown in Fig. 6, these include the water level signal (P<sub>v1\_WLS</sub>), smoke signal (P<sub>v3\_SKS</sub>), temperature control signal (P<sub>v2\_TCS</sub>), the motion detection signal (P<sub>v4\_MDS</sub>), and the keypad signal (P<sub>v5\_KPS</sub>). Whenever any of the qualifiers changes its state, either from logic 1 to logic 0 or from logic 0 to logic 1; there is an event at the output interface. For instance, in link path L1 to L4, when the input qualifier, water level signal (WLS) changes its state from logic 0 to logic 1 (L1-L2), Relay1 (the relay that turns on the pumping

machine) is energized and the water pump turns on. As long as it stays on in 1, the water pump will keep on pumping water. But once it changes its state from 1 to 0 (L3-L4), Relay1 is de-energized and the water pump turns off.

The link path L5 to L8 depicts what happens in the temperature channel and how the microcontroller reacts to it. When the temperature exceeds the preset value or goes below the preset value, the signal from the temperature module changes from logic 0 to logic1 (L5-L6), Relay 2 is energized and the air conditioner turns on. On the other hand, when the signal goes from 1 to 0 (L7-L8), the relay is de-energized and the air conditioner turns off. The L9-L12 depicts what happens in the smoke channel how the microcontroller reacts to it and what happens at the output interface. When there is a smoke occurrence, the signal from the smoke channel (one of the input qualifiers), changes from logic 0 to 1 L9-L10, the Buzzer is energized and turned on. For as long as the signal remains in logic 1, the buzzer will be sounding, but when a signal goes from logic 1 to 0 (L11-L12), the buzzer is de-energized and therefore stops sounding.

The link path L13-L16 depicts what happens in the motion-controlled light module and how the microcontroller reacts to it. When motion is detected and there is insufficient light, the signal from this module changes from logic 0 to 1, L13-L14, the relay known as RL3 is energized and the light turns on. As long as the person is in the room, the light will be on. But once the person leaves, that is, when the person is no longer sensed, the signal changes from logic 1 to 0, L15-L16 RL3 is de-energized and the light turns off.

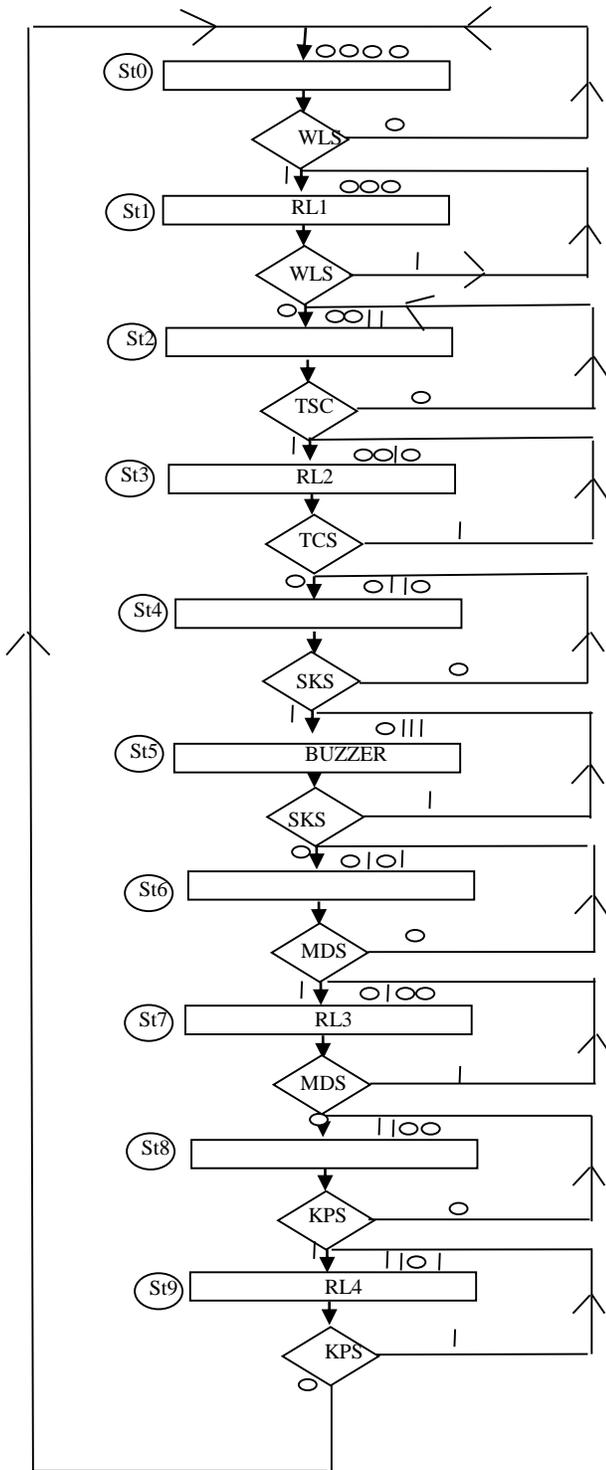


Fig. 6. MRASM Design Chart for Smart Automation and Security Design System.

The link path L17-L20 depicts what happens when the keypad is pressed. When pressed and the input code is correct, L17-L18, the signal from here changes from logic 0 to 1, the relay known as RL4 is energized and the door opens. Once the door is shut, the signal changes from logic 1 to 0, L19-L20. With the MRASM Chart and the corresponding, STT table, the system, home automation, and security system were developed

in the prototype design. The keypad signal (P<sub>v5\_KPS</sub>) goes into the PCon through port 3 pins 1 to 8 of the controller chip. The motion detection signal (P<sub>v4\_MDS</sub>) goes into the PCon through port 3 pins 12, 13, 14. The water level signal (P<sub>v1\_WLS</sub>) goes into the PCon through port 3 pins 15 and 16. The temperature control signal (P<sub>v2\_TCS</sub>,) goes into the PCon through pins 17. While the smoke signal (P<sub>v3\_SKS</sub>) goes in through pins 22 and 23. Fig. 7 shows the neuromorphic/schematic implementation described above.

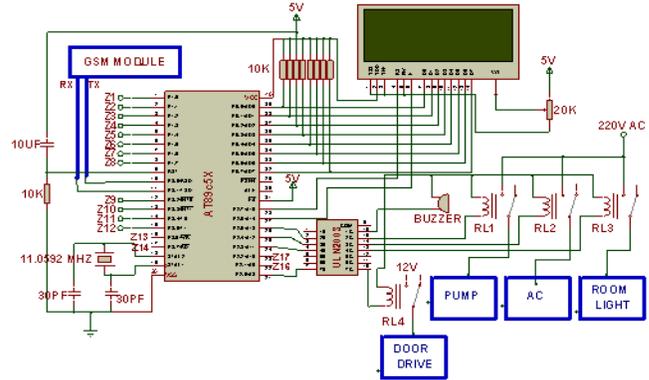


Fig. 7. MRASM-ROBOT Unified Schematics.

## V. FEEDBACK SCHEDULING ALGORITHM

### A. Dynamic Stability Control

Dynamic stability was introduced using digital real-time scheduler architecture in Fig. 8. This is implemented in Algorithm I, and II. The scheduler links the internal feedback control loop structures synchronously. Also, it sets the error margin for the scheduler, observes the error deviation state, and dynamically adjusts the infinite process variables for stable optimization. Algorithm II shows the temperature optimization algorithm. The AC, lights, pumps, and other processes are continuously monitored for efficient power management. At each instance, a control API communication is triggered for remote transmission.

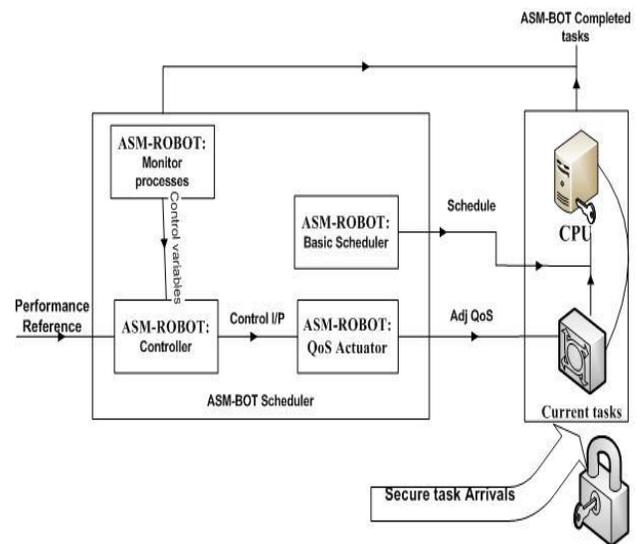


Fig. 8. MRASM-BOT Scheduler Architecture.

Internally, the deadline, estimated execution time, actual execution time, and baseline tasks are processed once the performance reference is set via initialization procedures. The MRASM-BOT captures the infinite process variables in the system and feeds samples to the controller. The controller makes comparisons between the performance reference and the controlled variables to ascertain immediate errors, and compute any error differential. The QoS actuator smartly adjusts the computed utilization at each sample scenario. The algorithm I demonstrate the optimization controls.

---

**Algorithm I.** Scheduling\_Optimization

**Input:** Set PID Gains (x, y, z) // gain service provisioning  
Set feedback static scheduling

**Output:** Keypad, Smoke, Temperature, Room\_light  
Set Pcon ON() // Set controller\_Monitor Active  
*recycle* every *monitorCallPeriod* minutes  
Initialize LCD\_pv1  
Initialize Modem\_pv2  
Initialize keypad\_pv3  
Initialize smoke sensor\_pv4  
Initialize Temperature\_pv5  
Initialize Room light\_pv6  
Initialize Water\_pv7

*Do forever*  
**if** (keypad\_pv2) **then**  
    Process (keypad) ← Predictor (*history* of Response Time)  
**Elseif** (smoke sensor)-Aware **then**  
    Process (Smoke\_parameterStatus) ← Predictor (*history* of parameterStatus)  
**end if**  
// CPS Planner  
**Else if** (Temperature) **then**  
    Process Temperature  
**Else if** (Room light) **then**  
    Process Room light  
**Else if** (water) **then**  
    Process (water level)  
End  
**Keypad:** Check for code  
**If** code parameter status is correct **then**  
    decision ← Grant\_Access  
**Elseif** parameterStatus allow “3” time check  
    decision ← Deny Access after #3  
**If** code parameterStatus is incorrect **then**  
    Decision ← Deny Access  
Send message  
Sound an alarm  
Display error (LCD)  
**End**  
**Smoke Sensor:** Check for smoke  
**If** (smoke sensed) **then**  
    Sound an alarm  
    Display message (LCD)  
**End**  
**Temperature:** Check Temperature  
**If** (temperature) too high **then**  
    Switch on “AC”  
    Else switch off “AC”  
**End**  
**Room Light:** Check entrance  
**If** (entrance) **then**

    Check room light intensity  
    **If** room dark **then**  
        Switch on light  
        Increment count  
    **Else** switch off light  
    **Else if** exit **then**  
        Check **if** count is “zero”  
    **If** not zero **then**  
        Decrement count  
        Switch off light  
    **End**  
    **Water level:** Check water level  
    **If** level is minimum **then**  
        Switch on pump  
        Check for maximum level  
    **If** level is maximum **then**  
        Display tank full  
        Switch off pump  
    **Elseif** decision is Power\_off **then**  
        | Excess\_PCon ← Don’t\_care\_state;  
    **end if**  
**end return**

---

In the neuromorphic automation model, once the message is sent, the message is received by SMSC (SMS controller). This then reaches an appropriate device/interface. Recall that the process controller (*PCon*) interface for SMS in Algorithm I provides a path for transmitting control signals in full-duplex mode.

A simple algorithm for IoT-SMS communication in the smart automation model is given in Algorithm II.

---

**Algorithm II.** IoT SMS Event Stream

**Input:** Set PID Gains (x, y, z) // gain service provisioning  
Set feedback static scheduling

**Output:** Keypad, Smoke, Temperature, Roomlight, etc  
Set Pcon ON() // Set controller\_Monitor Active

*Do forever*  
**Begin** (  
Step 1: Start  
Step 2: Process Controller Event initialization  
Step 3: Get Hardware Software  
Step 4: Poll SMS from AT Command for the Event  
Step 5: **If** abnormal condition at access layer sensor node, **then** go to step6 else, go to step1  
Step 6: Send SMS to mobile phone < Notify end-user >  
Step 7: Decode && Receive SMS based on Low BER  
Step 7: Check SMS pattern  
Step 8: Check Control device for own status  
Step 9: Take Corrective Action  
Step 10: Go to step1  
**Endif**  
**End**  
**end return**

---

**B. Reconfigurable Associative-Memory**

Recall from Table I, the memristive neuromorphic scenario is derived completely for all the machine states. The neuron circuitries with associative-memory-based reconfigurable neuromorphic (AMBRN) structure are implemented. In this case, learning and forgetting with associative memory dynamically reconfigures the circuit schematics. For this MRASM design, the design structure and its use-case are shown in Fig. 9.

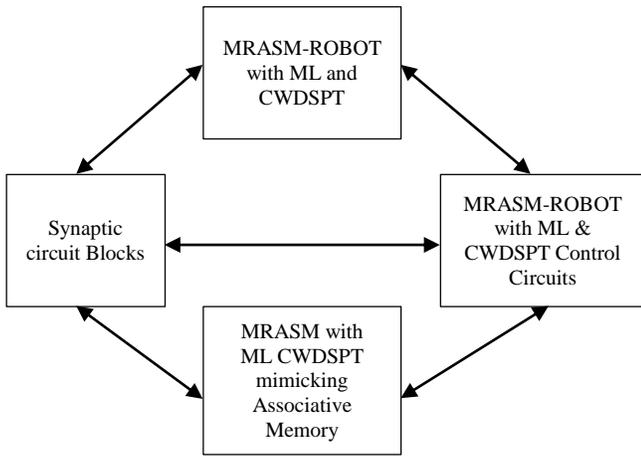


Fig. 9. Neuromorphic System Structure.

In Fig. 9, the AMBRN system has the following entities: 1) Machine learning subsystem block for speeding up process controls; 2) associative memory mimicking subsystem; 3) synaptic circuit block, 4) control circuits. The technique of mimicking the associative learning and the remembrance processes between conditioned and unconditioned process variable signals achieves the activation function. The remembrance process allows the control circuits to activate or deactivate links to the synaptic circuits. These synaptic circuits implement ML optimization in the robot, though this is currently under investigation. It realizes the associative memory network for dynamic responses.

## VI. SPACE DIVERSITY CONTROL MODEL

### A. IoT RF Modulation Construction

In this research, an investigation into the design properties of IoT RF transceiver was carried out in [66] leading to the model for computing error performance by measurements in Fig. 10. The model accounts for:

- Receiver sensitivity with different error measures (BER, Error SNR).
- Modulator frequency and phase error.
- Timing error deviation of the IoT RF transceiver.

Considering Fig. 10, the MRASM digital modulation process encodes data stream information from the sensed sources and makes it suitable for transmission. The modulation technique transports available data stream message/signal via radio channel. Best transmission quality on an optimal radio spectrum is leveraged.

Let's now, consider the SMS message from the robot as:

$$\psi(t) = \beta \cos(\omega t + \theta) \rightarrow \text{Channel} \quad (1)$$

Where  $\psi(t)$  = data streams,  $\beta \cos(\omega t + \theta)$  is the modulation. It is key to note that modulation is achieved via amplitude ( $A = \beta$ ) variation, phase  $\theta$ , high-frequency carrier ( $\omega t$ ) per the amplitude of the data streams signal. Also, Complexity Minimum Shift Keying (CMSK) is used for IoT RF optimization. The reason is that it offers a uniform envelope, optimal spectral efficiency, excellent bit error rate.

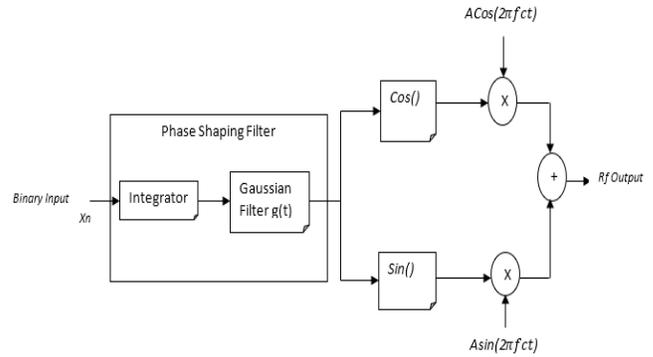


Fig. 10. MRASM IoT Modulation.

In context, the mathematical model for the modulated function  $X(t)$  is shown in (2).

$$x(t) = \text{Cos}(2\pi f_c + \phi(t)) \quad (2)$$

$$\phi(t) = 2\pi h \int_{-\infty}^{\infty} \sum_{k=-\infty}^{\infty} a_s s(\tau - kT) d\tau \quad (3)$$

### B. IoT GSMK Modulator

In the RF diversity design for MRASM-ROBOT in Fig. 10, complex IoT GSMK modulator/demodulator was used which has a Gaussian frequency shaping filter and key parameters from orthogonal space-time block code (OSTBC) transmission [66], sideband modulation [67], continuous phase modulation [68], GFSK Receiver [69], GSMK Pulse [70]. The scheme offers a continuous phase modulation (CPM) signal and has a modulation index  $h = 1/2$ . This means that the continuous phase shift function  $\theta(t)$  will have the complex baseband structure shown in (4).

$$\tau_{transmit}(t) = \beta \exp \sum_n^1 x_n \phi(t - nT) + \phi_0 \quad (4)$$

Where  $T$  is bit period,  $\beta$  is amplitude,  $X_n = \pm$  is the structural binary symbols,  $\phi_0$  is the random initial phase and  $\phi(t)$  is the phase shift function.

The unmodulated continuous-wave technique employed is robust and is not affected by signal fading and interference. Its spectral efficiency is optimal with slower/smooth phase changes. From Fig. 9, the IoT modulation signal is obtained through modulation and infusion of two quadrature carriers having frequency  $F_c$ . Phase changes are smoothed with a filter whose Gaussian impulse response is given by (5):

$$g(t) = \frac{1}{2T} Q 2 \left( \pi B \frac{t - \frac{T}{2}}{\sqrt{\ln 2}} \right) - Q \left( 2\pi B \frac{t + \frac{T}{2}}{\sqrt{\ln 2}} \right) \quad (5)$$

Where  $Q(t)$  is the Q-function given by (6):

$$Q(t) = \int_t^{\infty} \frac{1}{2\pi} \exp\left(-\frac{r^2}{2}\right) dr \quad (6)$$

and the phase shift function  $Q(t)$  in (1) is given by (7):

$$\phi(t) = \int_{-\infty}^t g(t) dt \quad (7)$$

In all, the IoT bandwidth and interference resistance parameters are controlled within space diversity by combination. The IoT modulation interface is shown in Fig. 11a. The source Bernoulli input is fed into the IoT RF

transmitter comprising convolutional encoding, data framing, interleaving, data burst, and cyclic redundancy checks (CRC). Also, its differential encoder with GMSK modulator is optimized for resilient transmission of process variables. The RF channel is constructed using additional white Gaussian noise (AWGN) and the Multipath Rayleigh fading. These were introduced for BER testing.

Similarly, the receiver the demodulation is derived from a decoder that has an isolated matched-filter, cyclic redundancy check, decoder, GMSK demodulation, differential-decoder, and reshape-optimizer. The receiver interface terminates with a BER/Error estimator sink.

In this work, BER is processed additive mapping, i.e., including source generated data with the demodulator output. Using GMSK modulation and demodulation processes in Fig. 11a leads to a stable but sensitive system. Form Fig. 10, the base BER is 0.03846 (i.e. less than 1). Hence, the system space diversity achieves absolute reliability as a transceiver unit.

In the MRASM-ROBOT, complex multiple-input multiple-output (MIMO) deployment especially in Rayleigh fading, can be determined.

Let's consider a complex IoT MIMO scenario with  $N_s$  Transmit antenna and  $N_d$  receive antennas, where  $N_s N_d \in \mathbb{Z}^+$ ,  $N_d > N_s$ , and  $N_d$  and  $N_s$  use huge values. The OSTB transmission for complex constellations in  $N_d * N_s$  MIMO system is derived from [66]. The MRASM-ROBOT channel matrix  $H \in \mathbb{C}^{N_d * N_s}$  can be computed from (8).

$$Y = HC + W \tag{8}$$

where  $C$  is the is code matrix, transmitted from transmitter and  $W$  is additive white Gaussian noise (AWGN) noise matrix.

By transforming (8) such that IoT OSTBC receiver combiner interface is shown in Fig. 11(b). The orthogonal code matrices  $C$ , decoder design are derived from [66]. The results for OSTBC are discussed later in Section VII.

Again, let's consider a single-IoT RF diversity design for which the signal received in (8) is the sum of the expected data stream signal and noise ( $W = n$ ) given by:

$$X = hu(t) + n \tag{9}$$

Where  $C = u(t)$  denotes IoT transmitted power signal.  $H = h$  denotes channel matrix with signal power.  $n$  the noise. The signal power sent out over a period,  $T_s$ , at nth element is given by (10).

$$P = \frac{1}{T_s} \int_0^{T_s} |h_n(t)|^2 |u(t)|^2 dt = |h_n(t)|^2 \frac{1}{T_s} \int_0^{T_s} |u(t)|^2 dt = |h_n(t)|^2, \tag{10}$$

Assuming slow fading due to small distance  $< 50m$ , the term  $|h_n(t)|^2$  will be kept constant for a period with an integral unit power  $E$ .

By letting  $E\{|h_n(t)|^2\} = \sigma^2$ , the instantaneous SNR at the  $n$ -th element ( $\gamma_n$ ) is given by (11).

$$\gamma_n = \frac{|h_n|^2}{\sigma^2} \tag{11}$$

The instantaneous SNR under channel matrix  $h_n$  is estimated such that noise power is obtained over a relatively short period  $T$ .

Given that Rayleigh fading is still possible over short-range, hence,  $h_n = [h_n]e^{j\angle h_n}$ , where  $\angle h_n$  is constant in  $(0, 2\pi)$ .  $[h_n]$  still has Rayleigh pdf, such that  $|h_n|^2$  and ( $\gamma_n$ ) has an exponential pdf given by (12).

$$[h_n] \sim \frac{2[h_n]}{P_0} e^{-|h_n|^2/P_0} \tag{12}$$

$$\gamma_n \sim \frac{1}{\Gamma} e^{-\gamma_n/\Gamma} \tag{13}$$

$$\Gamma = E\{\gamma_n\} = \frac{E\{|h_n|^2\}}{\sigma^2} = \frac{P_0}{\sigma^2} \tag{14}$$

Therefore, the instantaneous SNR at individual RF elements is exponentially distributed and  $r$  denotes the average SNR at each IoT element. This is also the SNR of a single element IoT RF antenna, i.e., the SNR with zero arrays. Hence,  $r$  is now the baseline for IoT RF SNR enhancement Fig. 11(b).

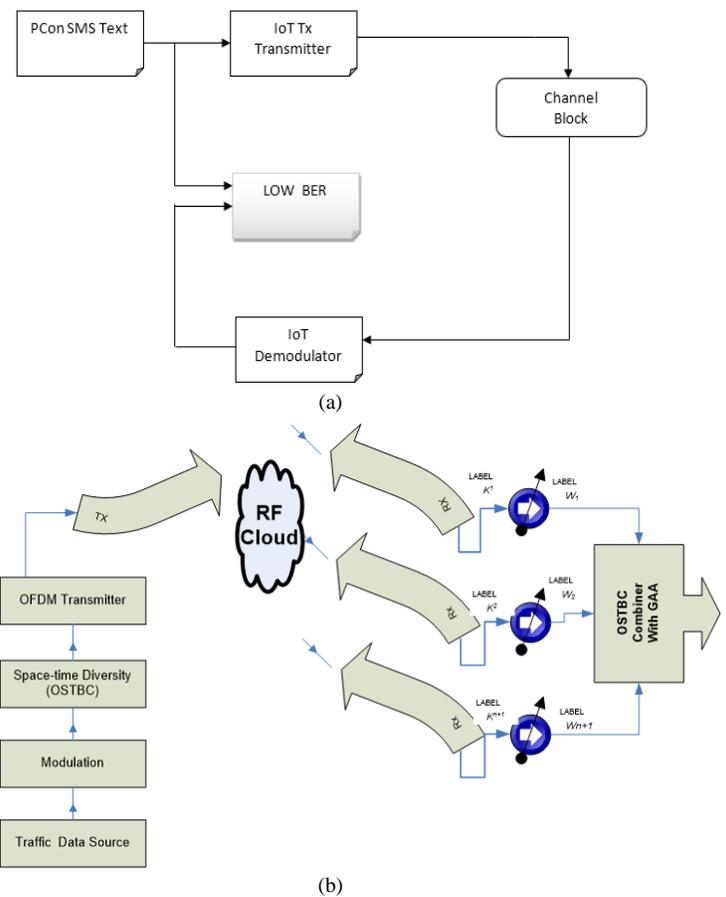


Fig. 11. (a) IoT Modulation Interface Model., (b). IoT OSTBC Receiver Combiner.

## VII. EXPERIMENTAL RESULTS

### A. MRASM-ROBOT Transmitter-Receiver Analysis

In this section, the experimental results to verify the proposed MRASM for the neuron control variables using space diversity theorems are presented. A brief discussion on the performance of the IoT CWDSPT scheme is highlighted. As a first step, C++ scripts was used to implement the control logic while MATLAB tool generated the plots. The implementation of the final version of the MRASM-ROBOT is currently undergoing packaging as a commercial off-the-shelf derivative.

For the IoT module, the receiver circuit is mapped at the carrier frequency  $FC = 2.4\text{GHz}$ . These schematics depict both embedded transmitter and receiver circuitry of the IoT module respectively. In Fig. 12a, the MRASM-ROBOT depicts the smart home automation and security system in an off-mode scenario. In Fig. 12b, the MRASM-ROBOT depicts the smart home automation and security system in the ON mode scenario.

### B. OSTBC Optimization Response

Recall in Section VI, the IoT GSMK Modulator was introduced. For process variables, the CWDSPT are affected by Multipath fading effects especially IoT RF interface that works with channel interference and AWGN. This paper introduced an enhanced layer 2 protocol called OSTBC. It is based on orthogonal frequency division multiplexing (OFDM) to suppress and enhance traffic frame delivery in the MRASM design. The OSTBC strategy in the receiver diversity block regulates efficient throughput by suppressing channel effects and recovering the transmitted signals.

Fig. 12 shows the error low pass filtering effect for noise reduction at receiving node while Fig. 13 shows the declining probability bit error rate (BER) under channel noise effect. These plots show the optimal capacity of the IoT CWDSPT scheme. This is because signal transmission and reception from other modules are achieved seamlessly.

The implication is that the MRASM-Robot supports process variable manipulation while reducing both the RF interference level (MRCSV) and (AWGN) for optimal performance.

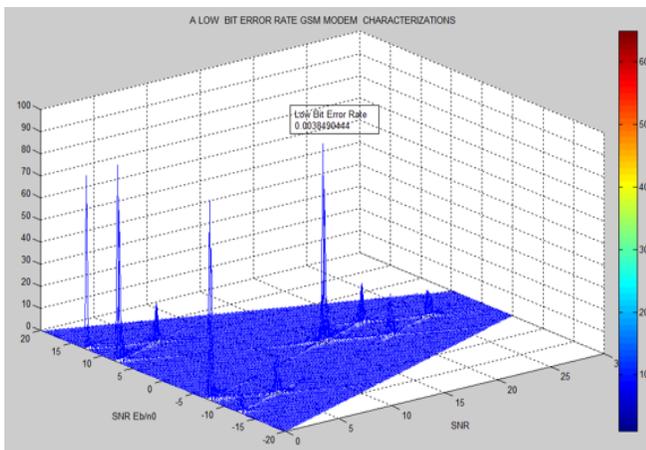


Fig. 12. MRASM-ROBOT Error Deviation with IoT Nodes.

Based on data generated, the relationship model using the modulation technique for BER is depicted in Table II. BER graph was obtained considering the AWGN channel and multipath fading. The GSMK modulation technique in the AWGN channel has good performance as shown in Fig. 12. When the robot was placed in the Multipath Rayleigh channel, an increasing value of Doppler shift (Hz) is shown to be acceptable. The implication is that the system will perform poorly as the coverage distance of the robot RF terminal is increased. Moreover, the system performs average well since the BER is quite low for such short-range communication distance. The results for BER vs. SNR are summarized in Table III.

TABLE II. SIMULATION SPECIFICATION [71]

| Simulation Parameters    | Specifications                     |
|--------------------------|------------------------------------|
| Scenario                 | Macro-Cell, 7-Nodes B's (21Sector) |
| Bs-2-Bs Distance         | 2800m (Large), 100m & 10m          |
| Cell Radius              | 933m                               |
| Propagation Model        | OSTBC Space diversity              |
| Channel Profile          | Multipath fading +AWGN             |
| Modulation               | OFDM bank                          |
| Channel bandwidth        | 3.5MHz                             |
| OFDM Symbols per burst   | 2                                  |
| Cyclic Prefix Factor (G) | 1/8                                |
| Receiver Type            | OFDM Receiver                      |
| Max.Doppler Shift (Hz)   | 0.5                                |
| Gain vector (db)         | [0 -5-10]                          |
| Initial Seed             | 61                                 |
| Channel SNR              | 20                                 |
| Number of Simulations    | 100                                |
| Traffic Model            | Infinite Buffer                    |

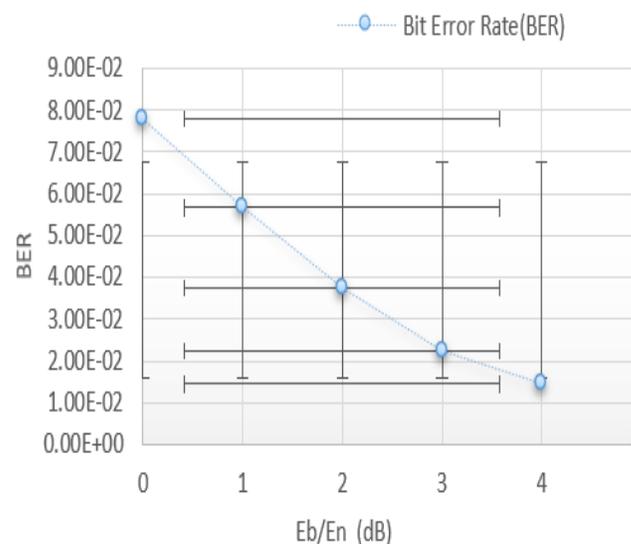


Fig. 13. Response Plot of MRASM-ROBOT under Channel Effects.

TABLE III. ERROR VALIDATION RESULTS

| SNR/E <sub>b</sub> /E <sub>0</sub> | Number of Errors | Bit Error Rate (BER) |
|------------------------------------|------------------|----------------------|
| 0                                  | 15615            | 7.81E-02             |
| 1                                  | 11334            | 5.67E-02             |
| 2                                  | 7520             | 3.76E-02             |
| 3                                  | 4481             | 2.24E-02             |
| 4                                  | 2489             | 1.44E-02             |

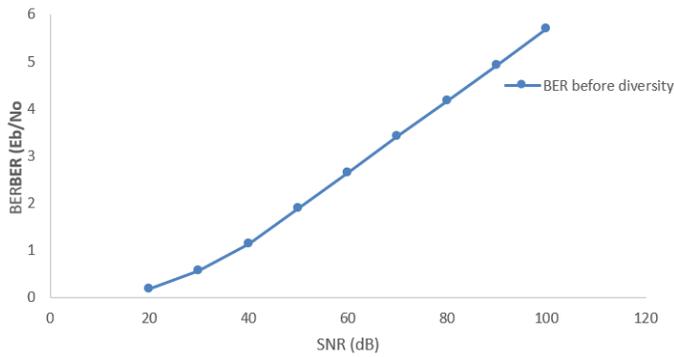


Fig. 14. MRASM-ROBOT BER before Diversity with IoT Nodes.

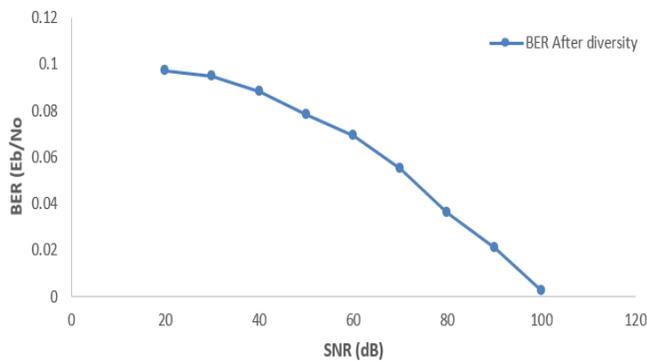


Fig. 15. MRASM-ROBOT BER after Diversity with IoT Nodes.

MRASM-ROBOT MIMO module is investigated for spatial diversity analysis since the robot communicates from a source to a sink always. The additive noise is varied with SNR. Space diversity analysis was carried out in Fig. 14 depicting ER before diversity and Fig. 1 denoting BER after diversity. Fig. 15 increases the reliability of the MRASM-ROBOT MIMO module. A second MRASM-ROBOT MIMO module (i.e., the IoT diversity antenna) below the first (i.e., the primary antenna) at each location of the MRASM-ROBOT having MIMO module will increase reliable connectivity.

Fig. 16, 17, 18, 19, and 20 shows that with the space diversity, the model will scale gracefully while offering better performance than the multipath fading schemes.

Under concurrent wireless data streams and power transfer, various space diversity simulation runs were executed and obtained polar graphs with different scenarios of CWDSPT (optimized and un-optimized). The analysis considered multipath channel fading effects in IoT RF interfaces. In the receiver diversity, with the channel interruptions, the plot no

OSTBC combiner is shown in Fig. 18 and 20. This leads to multipath fading channel issues and there is no memory stabilization since the OSTBC combiner block is absent. Fig. 19 and 21 depict highly optimized memory stabilization for the MRASM-ROBOT process variables. The optimization scenarios reflect the parametric sensitivities highlighted in Table III.

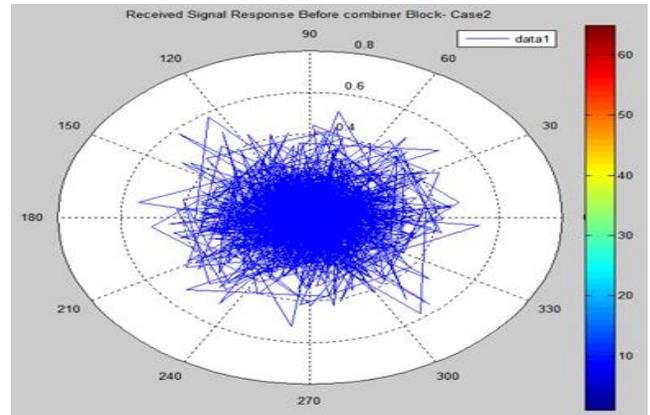


Fig. 16. Polar Plot of Received Signal without OSTBC Combiner Block (Case-2).

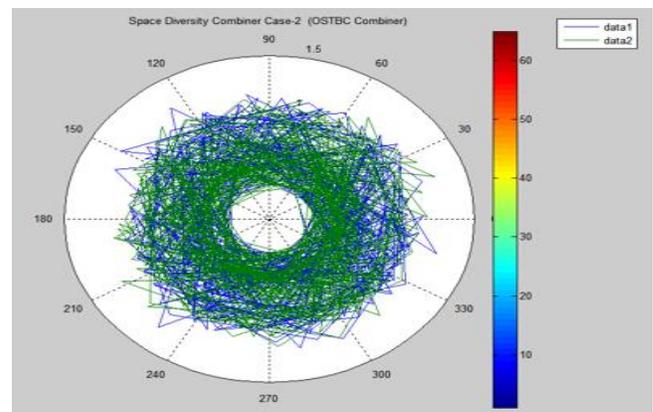


Fig. 17. Polar Plot of Received Signal with OSTBC Combiner Block, Case-2.

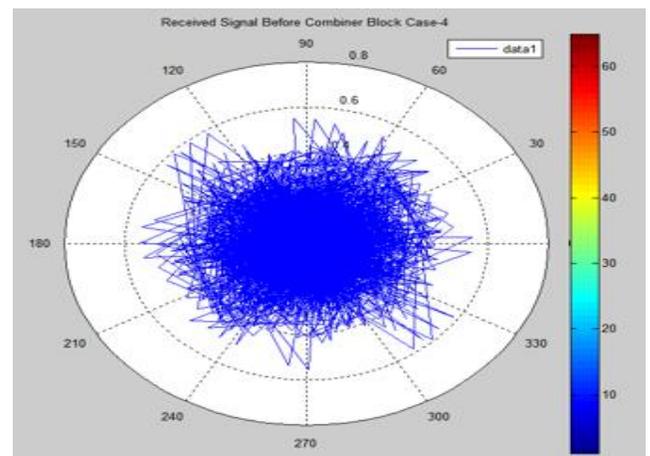


Fig. 18. Received Signal without Combiner Block, (Space Diversity Combiner Case -4).



TABLE IV. DIVERSITY COMPARISON

| Schemes        | BER        | SNR   | Reconfigurability |
|----------------|------------|-------|-------------------|
| HRSM-STBC [21] | $10^{-5}$  | 5.7dB | High              |
| OSTBC [66]     | $10^{-10}$ | 15dB  | Moderate          |
| MRASM-ROBOT    | $9^{-2}$   | 4dB   | Very High         |

### VIII. CONCLUSION

This paper has presented MRASM-ROBOT as a smart home automation system for both security challenged and process-driven environments. The reconfigurable memristive control strategy was used in the synaptic schematics. Dynamic stability with an error-free feedback control loop was realized for design architecture. The CWDSPT was introduced and tested with the OSTBC-CB space diversity scheme. For the process variables, control signals are transmitted with unmodulated high-power continuous wave (CW) for interference minimization. The design offered lower bit error rates leading to minimal error deviation for short distances.

The integration of IoT transmitter and receiver circuit for the device-to-device communication was implemented considering space diversity link reliability. Hence, CWDSPT signaling was optimized for telemetry data decoding within the deployment domain. The work showed the optimization polar plots with the OSTBC-CB for memory stabilization.

Future work will focus on the use of FPGA, containerization, machine learning, and cloud provisioning to address massive scalability concerns under RF antenna diversity. Also, process variable automation such as light systems, air conditioners, overhead tanks, security doors, among others, will be managed with containers.

### REFERENCES

[1] Y. Lv, Y. Fang, W. Chi, G. Chen, and L. Sun, "Object Detection for Sweeping Robots in Home Scenes (ODSR-IHS): A Novel Benchmark Dataset," *IEEE Access*, vol. 9, pp. 17820-17828, 2021.

[2] S. H. M. S. Andrade, G. O. Content, L. B. Rodrigues, L. X. Lima, N. L. Vijaykumar and C. R. L. Francês, "A Smart Home Architecture for Smart Energy Consumption in a Residence With Multiple Users," *IEEE Access*, vol. 9, pp. 16807-16824, 2021.

[3] M. -C. Su, J. -H. Chen, A. M. Arifai, S. -Y. Tsai and H. -H. Wei, "Smart Living: An Interactive Control System for Household Appliances," *IEEE Access*, vol. 9, pp. 14897-14904, 2021.

[4] R. Heartfield, G. Loukas, A. Bezemskij and E. Panaousis, "Self-Configurable Cyber-Physical Intrusion Detection for Smart Homes Using Reinforcement Learning," *IEEE Trans on Information Forensics and Sec.*, vol. 16, pp. 1720-1735, 2021.

[5] D. Riboni and F. Murru, "Unsupervised Recognition of Multi-Resident Activities in Smart-Homes," *IEEE Access*, vol. 8, pp. 201985-201994, 2020.

[6] J. Saunders, D. S. Syrdal, K. L. Koay, N. Burke and K. Dautenhahn, "Teach Me—Show Me"—End-User Personalization of a Smart Home and Companion Robot," *IEEE Trans. on Human-Machine Systems*, 46(1), pp. 27-40, 2016.

[7] Y. Zhang and Z. Zeng, "A Multi-functional Memristive Pavlov Associative Memory Circuit Based on Neural Mechanisms," in *IEEE Transactions on Biomedical Circuits and Systems*, Pg 1-1, Early Access. doi: 10.1109/TBCAS.2021.3108354.

[8] S. Vahdat, M. Kamal, A. Afzali-Kusha and M. Pedram, "Reliability Enhancement of Inverter-Based Memristor Crossbar Neural Networks Using Mathematical Analysis of Circuit Non-Idealities," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 10, pp. 4310-4323, Oct. 2021, doi: 10.1109/TCSI.2021.3105043.

[9] Y. Song, Q. Wu, X. Wang, C. Wang and X. Miao, "Two Memristors-Based XOR Logic Demonstrated With Encryption/Decryption," in *IEEE Electron Device Letters*, 42(9), pp. 1398-1401, Sept. 2021, doi: 10.1109/LED.2021.3102678.

[10] X. Xu et al., "MDA: A Reconfigurable Memristor-Based Distance Accelerator for Time Series Mining on Data Centers," in *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Sys*, 38(5), pp. 785-797, 2019, doi: 10.1109/TCAD.2018.2834431.

[11] M. Nourazar, V. Rashtchi, A. Azarpeyvand and F. Merrikh-Bayat, "Code Acceleration Using Memristor-Based Approximate Matrix Multiplier: Application to Convolutional Neural Networks," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 26(12), pp. 2684-2695, Dec. 2018, doi: 10.1109/TVLSI.2018.2837908.

[12] X. Xu et al., "Accelerating Dynamic Time Warping With Memristor-Based Customized Fabrics," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(4), pp. 729-741, April 2018, doi: 10.1109/TCAD.2017.2729344.

[13] L. Yavits, R. Kaplan and R. Ginosar, "GIRAF: General Purpose In-Storage Resistive Associative Framework," 33(2), pp. 276-287, 1 Feb. 2022, doi: 10.1109/TPDS.2021.3065448.

[14] M. Edmonds, T. Atahary, S. Douglass and T. Taha, "Hardware Accelerated Semantic Declarative Memory Systems through CUDA and MapReduce," in *IEEE Transactions on Parallel and Distributed Systems*, 30(3), pp. 601-614, 1. 2019, doi: 10.1109/TPDS.2018.2866848.

[15] L. Chen, C. Li, X. Wang, and S. Duan, "Associate learning and correcting in a memristive neural network," *Neural Comput. Appl.*, vol. 22, no. 6, pp. 1071-1076, 2013.

[16] L. Yang and Z. Zeng, "A memristor-CMOS hybrid circuit for classical conditioning reflex," in *Proc. 18th Int. Conf. Inf. Sci. Technol.*, 2018, pp. 257-261.

[17] J. Sun, G. Han, Z. Zeng and Y. Wang, "Memristor-Based Neural Network Circuit of Full-Function Pavlov Associative Memory With Time Delay and Variable Learning Rate," in *IEEE Transactions on Cybernetics*, 50(7), pp.2935-2945, 2020,doi: 10.1109/TCYB.2019.2951520.

[18] M. Ansari et al., "PHAX: Physical Characteristics Aware Ex-Situ Training Framework for Inverter-Based Memristive Neuromorphic Circuits," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(8), pp. 1602-1613, Aug. 2018, doi: 10.1109/TCAD.2017.2764070.

[19] A. Hamdan, H. Hijazi, L. Ros, C. Siclet and A. Al-Ghouwayel, "Interference Analysis for Multi-Carrier Systems Over Fast-Fading Multipath Channels," *IEEE Latin-American Conference on Communications (LATINCOM)*, 2021, pp. 1-6, doi: 10.1109/LATINCOM53176.2021.9647847.

[20] C. Jeong and S. H. Chae, "Simultaneous Wireless Information and Power Transfer for Multiuser UAV-Enabled IoT Networks," in *IEEE Internet of Things Journal*, 8(10), pp. 8044-8055, 15, 2021, doi: 10.1109/JIOT.2020.3043210.

[21] X. N. Tran, X. Nguyen, M. Le and V. Ngo, "High-Rate Spatially Modulated Space Time Block Code," in *IEEE Communications Letters*, 22(12), pp. 2595-2598,2018, doi: 10.1109/LCOMM.2018.2872938.

[22] B. Clerckx, J. Kim, K. W. Choi and D. I. Kim, "Foundations of Wireless Information and Power Transfer: Theory, Prototypes, and Experiments," in *Proceedings of the IEEE*, vol. 110, no. 1, pp. 8-30, Jan. 2022, doi: 10.1109/JPROC.2021.3132369.

[23] I. A. Hernandez-Robles, X. Gonzalez-Ramirez, N. D. Galan-Hernandez and J. M. Ramirez-Arredondo, "Analysis and Design Tool for Wireless Power Transfer for Multiple Applications Purposes," in *IEEE Canadian Journal of Electrical and Computer Engineering*, 45(1), pp. 24-30, winter 2022, doi: 10.1109/ICJECE.2021.3099402.

[24] L. Wang, J. Li, H. Chen and Z. Pan, "Radial-Flux Rotational Wireless Power Transfer System With Rotor State Identification," in *IEEE Transactions on Power Electronics*, 37(5), pp. 6206-6216, May 2022, doi: 10.1109/TPEL.2021.3132702.

[25] J. Zhou, P. Zhang, J. Han, L. Li and Y. Huang, "Metamaterials and Metasurfaces for Wireless Power Transfer and Energy Harvesting," in *Proceedings of the IEEE*, 110(1), pp. 31-55, 2022, doi: 10.1109/JPROC.2021.3127493..

- [26] X. Wang, Z. Ning, L. Guo, S. Guo, X. Gao and G. Wang, "Online Learning for Distributed Computation Offloading in Wireless Powered Mobile Edge Computing Networks," in *IEEE Transactions on Parallel and Distributed Systems*, 33(8), pp. 1841-1855, 1 Aug. 2022, doi: 10.1109/TPDS.2021.3129618.
- [27] T. Ma, Y. Wang, X. Hu, D. Zhao, Y. Jiang and C. Jiang, "Periodic Energy Control for Wireless Power Transfer System," in *IEEE Transactions on Power Electronics*, 37(4), pp. 3775-3780, April 2022, doi: 10.1109/TPEL.2021.3129501.
- [28] Z. -R. Xu, Y. -F. Ye, L. -S. Wu and J. -F. Mao, "Microstrip Memristive Switch and Its Applications to RF Devices," 2020 International Conference on Microwave and Millimeter Wave Technology (ICMMT), 2020, pp. 1-3, doi: 10.1109/ICMMT49418.2020.9386898.
- [29] S. Yu, A. K. Das and Y. Park, "Comments on "ALAM: Anonymous Lightweight Authentication Mechanism for SDN Enabled Smart Homes"," in *IEEE Access*, vol. 9, pp. 49154-49159, 2021.
- [30] M. Ansari et al., "PHAX: Physical Characteristics Aware Ex-Situ Training Framework for Inverter-Based Memristive Neuromorphic Circuits," in *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(8), pp. 1602-1613, 2018, doi: 10.1109/TCAD.2017.2764070.
- [31] S. Vahdat, M. Kamal, A. Afzali-Kusha and M. Pedram, "Reliability Enhancement of Inverter-Based Memristor Crossbar Neural Networks Using Mathematical Analysis of Circuit Non-Idealities," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(10), pp. 4310-4323, Oct. 2021, doi: 10.1109/TCSI.2021.3105043.
- [32] Z. Fahimi, M. R. Mahmoodi, M. Klachko, H. Nili and D. B. Strukov, "The Impact of Device Uniformity on Functionality of Analog Passively-Integrated Memristive Circuits," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(10), pp. 4090-4101, 2021, doi: 10.1109/TCSI.2021.3097282.
- [33] M. I. Khan, S. Ali, A. A. Ikram and A. Bermak, "Optimization of Memristive Crossbar Array for Physical Unclonable Function Applications," in *IEEE Access*, vol. 9, pp. 84480-84489, 2021, doi: 10.1109/ACCESS.2021.3087810.
- [34] T. Titirsha et al., "Endurance-Aware Mapping of Spiking Neural Networks to Neuromorphic Hardware," in *IEEE Transactions on Parallel and Distributed Systems*, 33( 2), pp. 288-301, 1 2022, doi: 10.1109/TPDS.2021.3065591.
- [35] Y. Zhang and Z. Zeng, "A Multi-functional Memristive Pavlov Associative Memory Circuit Based on Neural Mechanisms," in *IEEE Transactions on Biomedical Circuits and Systems*, 15(5), pp. 978-993, 2021, doi: 10.1109/TBCAS.2021.3108354.
- [36] L. Wang, H. Li, S. Duan, T. Huang, and H. Wang, "Pavlov associative memory in a memristive neural network and its circuit implementation," *Neurocomputing*, vol. 171, pp. 23-29, Jan. 2016.
- [37] Y. Zhou, X. Hu, L. Wang, G. Zhou and S. Duan, "QuantBayes: Weight Optimization for Memristive Neural Networks via Quantization-Aware Bayesian Inference," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(12), pp.4851-4861, 2021, doi: 10.1109/TCSI.2021.3115787.
- [38] D. Liang, R. Kreiser, C. Nielsen, N. Qiao, Y. Sandamirskaya and G. Indiveri, "Neural State Machines for Robust Learning and Control of Neuromorphic Agents," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(4), pp. 679-689, 2019, doi: 10.1109/JETCAS.2019.2951442.
- [39] S. Roy, A. Banerjee and A. Basu, "Liquid State Machine With Dendritically Enhanced Readout for Low-Power, Neuromorphic VLSI Implementations," in *IEEE Transactions on Biomedical Circuits and Systems*, 8,(5), pp. 681-695,2014, doi: 10.1109/TBCAS.2014.2362969.
- [40] D. Liang and G. Indiveri, "A Neuromorphic Computational Primitive for Robust Context-Dependent Decision Making and Context-Dependent Stochastic Computation," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, 66,(5), pp. 843-847, 2019, doi: 10.1109/TCSI.2019.2907848.
- [41] L. Wang, H. Li, S. Duan, T. Huang, and H. Wang, "Pavlov associative memory in a memristive neural network and its circuit implementation," *Neurocomputing*, vol. 171, pp. 23-29, Jan. 2016.
- [42] Y. Wang, W. Fei, and H. Yu, "SPICE simulator for hybrid CMOS memristor circuit and system," in *Proc. Cellular Nanoscale Netw. Appl.*, 2012, pp. 1-6.
- [43] C. Mohan, L. A. Camuñas-Mesa, J. M. De La Rosa, E. Vianello, T. Serrano-Gotarredona and B. Linares-Barranco, "Neuromorphic Low-Power Inference on Memristive Crossbars With On-Chip Offset Calibration," in *IEEE Access*, vol. 9, pp. 38043-38061, 2021, doi: 10.1109/ACCESS.2021.3063437.
- [44] L. Yang, Z. Zeng and Y. Huang, "An Associative-Memory-Based Reconfigurable Memristive Neuromorphic System With Synchronous Weight Training," in *IEEE Trans on Cognitive and Developmental Syst.*, 12(3), pp. 529-540, 2020, doi: 10.1109/TCDS.2019.2932179.
- [45] 41H. An, Q. An and Y. Yi, "Realizing Behavior Level Associative Memory Learning Through Three-Dimensional Memristor-Based Neuromorphic Circuits," in *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(4), pp. 668-678, 2021. doi: 10.1109/TETCI.2019.2921787.
- [46] 42M. Payvand, M. E. Fouda, F. Kurdahi, A. M. Eltawil and E. O. Neftci, "On-Chip Error-Triggered Learning of Multi-Layer Memristive Spiking Neural Networks," in *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 10(4), pp. 522-535, 2020, doi: 10.1109/JETCAS.2020.3040248.
- [47] 43K. Bai, Q. An, L. Liu and Y. Yi, "A Training-Efficient Hybrid-Structured Deep Neural Network With Reconfigurable Memristive Synapses," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 28(1), pp. 62-75, Jan. 2020, doi: 10.1109/TVLSI.2019.2942267.
- [48] B. Setz, S. Graef, D. Ivanova, A. Tiessen and M. Aiello, "A Comparison of Open-Source Home Automation Systems," in *IEEE Access*, vol. 9, pp. 167332-167352, 2021, doi: 10.1109/ACCESS.2021.3136025.
- [49] P. Franco, J. M. Martínez, Y. -C. Kim and M. A. Ahmed, "A Framework for IoT Based Appliance Recognition in Smart Homes," in *IEEE Access*, vol. 9, pp. 133940-133960, 2021, doi: 10.1109/ACCESS.2021.3116148.
- [50] S. Yu, N. Jho and Y. Park, "Lightweight Three-Factor-Based Privacy-Preserving Authentication Scheme for IoT-Enabled Smart Homes," in *IEEE Access*, vol. 9, pp. 126186-126197, 2021, doi: 10.1109/ACCESS.2021.3111443.
- [51] M. J. Iqbal et al., "Smart Home Automation Using Intelligent Electricity Dispatch," in *IEEE Access*, vol. 9, pp. 118077-118086, 2021, doi: 10.1109/ACCESS.2021.3106541.
- [52] X. Ran and S. Leng, "Enhanced Robust Index Model for Load Scheduling of a Home Energy Local Network With a Load Shifting Strategy," in *IEEE Access*, vol. 7, pp. 19943-19953, 2019, doi: 10.1109/ACCESS.2018.2889762.
- [53] D. Lan, Z. Pang, C. Fischione, Y. Liu, A. Taherkordi and F. Eliassen, "Latency Analysis of Wireless Networks for Proximity Services in Smart Home and Building Automation: The Case of Thread," in *IEEE Access*, vol. 7, pp. 4856-4867, 2019, doi: 10.1109/ACCESS.2018.2888939.
- [54] R. Zhang et al., "An EOG-Based Human-Machine Interface to Control a Smart Home Environment for Patients With Severe Spinal Cord Injuries," in *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 1, pp. 89-100, 2019, doi: 10.1109/TBME.2018.2834555.
- [55] P. N. Dawadi, D. J. Cook and M. Schmitter-Edgecombe, "Automated Cognitive Health Assessment Using Smart Home Monitoring of Complex Tasks," in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 43(6), pp. 1302-1313, 2013, doi: 10.1109/TSMC.2013.2252338.
- [56] S. D. T. Kelly, N. K. Suryadevara and S. C. Mukhopadhyay, "Towards the Implementation of IoT for Environmental Condition Monitoring in Homes," in *IEEE Sensors Journal*, 13(10), pp. 3846-3853, Oct. 2013, doi: 10.1109/JSEN.2013.2263379.
- [57] J. -m. Choi, B. -k. Ahn, Y. -s. Cha and T. -y. Kuc, "Remote-controlled Home Robot Server with Zigbee Sensor Network," *SICE-ICASE International Joint Conference*, 2006, pp. 3739-3743, doi: 10.1109/SICE.2006.315025.
- [58] J. Saunders, D. S. Syrdal, K. L. Koay, N. Burke and K. Dautenhahn, "'Teach Me-Show Me"—End-User Personalization of a Smart Home

- and Companion Robot," in *IEEE Transactions on Human-Machine Systems*, 46(1), pp. 27-40, 2016, doi: 10.1109/THMS.2015.2445105.
- [59] U. Kim and J. Kim, "A Stabilized Feedback Episodic Memory (SF-EM) and Home Service Provision Framework for Robot and IoT Collaboration," in *IEEE Transactions on Cybernetics*, 50(5), pp. 2110-2123, May 2020, doi: 10.1109/TCYB.2018.2882921.
- [60] K. Park, H. Lee, Y. Kim and Z. Z. Bien, "A Steward Robot for Human-Friendly Human-Machine Interaction in a Smart House Environment," in *IEEE Transactions on Automation Science and Engineering*, 5(1), pp. 21-25, Jan. 2008, doi: 10.1109/TASE.2007.911674.
- [61] A. Corrales Paredes, M. Malfaz and M. A. Salichs, "Signage System for the Navigation of Autonomous Robots in Indoor Environments," in *IEEE Transactions on Industrial Informatics*, 10(1), pp. 680-688, Feb. 2014, doi: 10.1109/TII.2013.2246173.
- [62] J. Berrezueta-Guzman, I. Pau, M. -L. Martín-Ruiz and N. Máximo-Bocanegra, "Smart-Home Environment to Support Homework Activities for Children," in *IEEE Access*, vol. 8, pp. 160251-160267, 2020, doi: 10.1109/ACCESS.2020.3020734.
- [63] C. Messeri, A. Bicchi, A. M. Zanchettin and P. Rocco, "A Dynamic Task Allocation Strategy to Mitigate the Human Physical Fatigue in Collaborative Robotics," in *IEEE Robotics and Automation Letters*, 7(2), pp. 2178-2185, 2022, doi: 10.1109/LRA.2022.3143520.
- [64] H. Tang, A. Wang, F. Xue, J. Yang and Y. Cao, "A Novel Hierarchical Soft Actor-Critic Algorithm for Multi-Logistics Robots Task Allocation," in *IEEE Access*, vol. 9, pp. 42568-42582, 2021, doi: 10.1109/ACCESS.2021.3062457.
- [65] C. Mohan, L. A. Camuñas-Mesa, J. M. De La Rosa, E. Vianello, T. Serrano-Gotarredona and B. Linares-Barranco, "Neuromorphic Low-Power Inference on Memristive Crossbars With On-Chip Offset Calibration," in *IEEE Access*, vol. 9, pp. 38043-38061, 2021, doi: 10.1109/ACCESS.2021.3063437.
- [66] A. M. K., "OSTBC Transmission in Large MIMO Systems," in *IEEE Communications Letters*, 20(11), pp. 2308-2311, 2016, doi: 10.1109/LCOMM.2016.2597229.
- [67] K. Kassar, H. Farès, D. C. Glattli and Y. Louët, "Performance vs. Spectral Properties for Single-Sideband Continuous Phase Modulation," in *IEEE Trans on Comm*, 69(7), pp. 4402-4416, 2021, doi: 10.1109/TCOMM.2021.3073792.
- [68] Y. Sun, "Optimal Parameter Design of Continuous Phase Modulation for Future GNSS Signals," in *IEEE Access*, vol. 9, pp. 58487-58502, 2021, doi: 10.1109/ACCESS.2021.3073317.
- [69] J. Zhao, Y. Zhang, K. Zeng, W. Rhee and Z. Wang, "A 2.4-GHz Crystal-Less GFSK Receiver Using an Auxiliary Multiphase BBPLL for Digital Output Demodulation With Enhanced Frequency Scaling," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no.4, pp. 1143-1147, 2021, doi: 10.1109/TCSII.2020.3032149.
- [70] R. Ahmad and A. Srivastava, "PAPR Reduction of OFDM Signal Through DFT Precoding and GMSK Pulse Shaping in Indoor VLC," in *IEEE Access*, vol. 8, pp. 122092-122103, 2020, doi: 10.1109/ACCESS.2020.3006247.
- [71] K. C. Okafor; Guinevere, E.C.; Akinyele, O.O, "Hardware Description Language (HDL): An Efficient Approach to Device Independent Designs for VLSI Market Segments", *IEEE Int'l Conf Adaptive Science and Technology (ICAST)*, 2011, Abuja, 24th-26th, 2011. Pp. 262 – 267. DOI: 10.1109/ICASTech.2011.6145181
- [72] K. C. Okafor, G.C. Ononiwu, Sam G. V.C Chijindu, C. C. Udeze "Towards Complex Dynamic Fog Network Orchestration Using Embedded Neural Switch", *In International Journal of Computers and Applications, (IJCA), UK.*, 2021, 43(2), Pp.91-108.